

Determining Historic All-NBA Teams Using a Two Backcourt - Three Frontcourt Format with Machine Learning

Jeremy Lu

1. Abstract

This paper revisits historic All-NBA teams, and aims to predict . The All-NBA team is voted upon by the media in five man lineups, meant to highlight the best players in the NBA regular season. Traditionally, All-NBA teams features two guards, two forwards, and a center; but with the modern NBA, the lines between positions have been blurred, with players taking on a variety of roles that may be deemed “non-traditional”. To cater to the era of positionless basketball, this paper reimagines All-NBA team consisting of two backcourt (guards) and three frontcourt players (forward OR center). Training a linear regression and random forest regressor model on voting data since the 1988-89 season, this paper then predicts the All-NBA teams for the 2021-22 season, as well as analyzes changes to historic All-NBA lineups as well.

2. All-NBA Background

2.1. History

The All-NBA team is an annual award voted upon by a group of global sportswriters and broadcasters that recognizes fifteen total players for exceptional regular season performance. Voters submit three five man lineups, consisting of two guards, two forwards, and center. Players on the first team receive five points, second team receive three, and third team receive one. Then, the points are tallied and First, Second, and Third team members are selected from the point totals. From 1946-47 to 1954-55, a first and second team was selected without regard to position, but positions were added with two guards, two forwards, and a center in the 1955-56 season. The next change was made before the 1988-89 season, with the addition of a third team.

Typically, players are selected to the All-NBA team based on individual performance, with preference given to scoring. However, other factors such as team success, games played, etc... are often considered as well. A lack of clear criteria provided by the league makes this award quite subjective, with discourse happening frequently regarding voting results.

2.2. Position-less Basketball

As players have become more skilled and athletic, traditional positions and their roles have become outdated. Point guards, the primary playmakers who are usually short statured, have evolved to include super sized ball handlers like Luka Doncic, who is listed as 6'7, the 6'10 Ben Simmons, or newcomer Cade Cunningham, who measures in at 6'6. Shooting guards and small forwards have really merged into one position, essentially wing players with a wide variety of skills to support the point guard on the perimeter. Meanwhile, the traditional power forward has a diminished role; with shooting becoming a premium skill, forwards without floor spacing ability find it much harder to stay in the league. Finally, centers, who used to be lumbering big men that dominated the paint, have expanded their game to all areas of the court, with centers like Joel Embiid and Nikola Jokic shooting three pointers and taking stepback jumpers. While players that fit traditional positional moulds still exist, the increased skill demand from players across all positions has changed the landscape of the league as well as forced us to reimagine player positions.

This year has been especially unique given the landscape of the league, with teams have taken “position-less” basketball to the extreme. For example, the Boston Celtics and Toronto Raptors both start 4 players ranging from the height of 6'6 to 6'9, opting out of traditional centers. The NBA also made Jokic and Embiid eligible at both forward AND center (seemingly to allow for an Antetokounmpo, Jokic, Embiid pairing in the All NBA First Team), which drew mixed reactions from fans and media. While many voters claimed they would stick to one “traditional” center in their ballot, others celebrated this change, citing how the All Star game voting follows the two backcourt-three frontcourt format as well.

In several past occasions, certain players would stand to benefit from a changed All-NBA ballot. In 2015, Deandre Jordan was selected to All-NBA First Team as a center, in a year

he averaged just 12.7 points and 13.8 rebounds a game. Meanwhile, Draymond Green, who had 14 points, 9.5 rebounds, and 7.4 assists, had more total voting points than Jordan; but because those points were split between forward and center, lost out on First Team status. More recently in 2020, Khris Middleton received more total points than Ben Simmons and Russell Westbrook, but because of positional technicalities, was not named to any All NBA team.

Taking a step back in time to the 1993-94 season, Patrick Ewing did not make the All-NBA team, as he was behind Hakeem Olajuwon, David Robinson, and Shaquille O’Neal in the center spots. Yet he finished 5th in MVP voting ahead of players like Shawn Kemp and Karl Malone who were named to All-NBA.

2.3. Summary

Overall, as basketball has trended to become “position-less”, there has been a growing support for a best 15 player set up for the All-NBA teams as opposed to the current two guard, two forward, one center setup. While throwing positions entirely out of the conversation is quite radical, the two backcourt-three frontcourt setup currently used in NBA All Star voting is a viable alternative. As seen in previous examples, positional technicalities can and do affect results, and such a change would better reflect the modern landscape of the NBA. This paper uses this option to explore the ramifications of such a change both for present and past All NBA voting.

3. Data Collection and Processing

3.1. Data Collection

From [Basketball Reference](#), we scrape NBA Award Voting, Regular Season Standings, and Regular Season Player Stats Per 100 Possessions from the 1988-89 season to the 2021-22 season. 1988-89 is the chosen start season since this is when an All-NBA Third Team was chosen.

For regular season player stats, we take the basic counting stats: games, points, assists, rebounds, steals, blocks, and turnovers. Additionally, we have basic shooting splits: field goal, three point, and free throw percentage as well as the number of attempts per possession

of each. In total, we have 19095 rows of data across 34 seasons. An example page of the data can be accessed [here](#).

For standings, we simply collect the team record of each team for all our seasons, as well as winning percent, and the rank. Rank is important here as we often reference teams as being “the best (or n -th best) in the league” based on record. Year to year variation means the best performing team will have a different winning percentage, so rank is more useful to illustrate team performance in comparison to the rest of the league. As an example the 2015-16 Spurs were only the second best team with a record of 67-15, while the SuperSonics had the best record at 63-19 during the 1993-94 season; here rank would better measure team performance, as looking purely at wins may suggest the Spurs were the “better” team. In total, we have 988 teams across 34 seasons, with all of them being 82 game seasons with the exception of 1998-1999 and 2011-12 seasons (lockout shortened to 50 and 66 games, respectively). An example page of the data can be accessed [here](#).

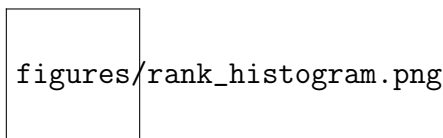


Figure 1: Here we can see that players selected to All-NBA teams are usually on better teams

Finally, for All-NBA voting data, we collect the player, their position and team, as well as the number of voting points they receive and vote share (percentage of total voting points received). Because the number of voters is not consistent across seasons, we aim to use vote share as the outcome variable. This will also allow us to use more models as our outcome is not just binary. Finally, vote share is also a better metric as there are often close races for All NBA spots, and simply a “yes” or “no” prediction fails to capture that perhaps there are several players deserving consideration for All NBA honors. Across all 33 seasons (no 2021-22 results yet), we have 1369 total players receiving votes, and a total of 495 All NBA spots. An example page of the data collected can be viewed [here](#) in the All-NBA Teams section.

3.2. Data Processing

Upon collecting three tables of data, there are several steps taken to prepare our training and test data, as long as the rationale behind these decisions:

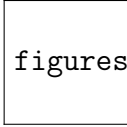
1. The minimum percentage of games played from a player that made an All-NBA team is 0.5365854. Thus an arbitrary cutoff of 0.5 was used for the minimum percentage of games that players had to participate in to count in the data.

Establishing a cutoff is essential because players that perform well but simply do not participate in enough games are not as valuable in the context of a season. Additionally, some bench players who make a single shot in a garbage time game would have a field goal percentage of 1, and thus the numbers would indicate that he is the best shooter in the league (this is clearly not true); we want to remove statistical outliers like these. This step resulted in keeping 11484 out of the 19095 rows of player data collected.

2. Some player names have characters with non UTF-8 recognized characters. Using the `replace_non_ascii` function from the `textclean` package, we can turn these characters into a recognizable format for R. Now, we can join our player data and voting data based on player and year, and this gives us 1369 rows.
3. At this point, there are certain players whose team is “TOT”, indicating they were traded midseason. Since a player traded from Team A to Team B will have contributed to Team B, their team will be set from “TOT” to the abbreviation of the receiving team. For example, during the 2020-21 season, James Harden was traded from the Houston Rockets to the Brooklyn Nets. In our data, he will be designated as a member of the Nets. This results in 1352 entries in our data.

4. After this step, we group our data by year and standardize all counting stats.

The grouping by year is done because All-NBA selections are single season awards made in comparison to other players across that one season; not across other seasons. Thus standardizing our data will allow us to measure how much better a player is compared to his peers in that season. With the pace of the game increasing and rules opening offenses, scoring rates have increased, as shown here:



figures/points_bar.png

Figure 2: The amount of players that score 30+ points per 100 possessions has increased over time

Overall, standardizing allows us to make within season comparisons better, and we apply this towards all counting (and shooting) stats.

5. Next, we join our data with the standings data, which adds team performance into our dataframe. This will allow the model to factor in team performance when predicting vote share for All-NBA. A common argument against skilled up and coming players on bad teams (think Kings DeMarcus Cousins, Devin Booker pre-CP3, or Wolves Zach LaVine) is that they are “empty-calorie” players, guys that score because their team is full of nobodies and who fail to elevate their team or play “winning basketball”. It will be interesting to see the impact of this on players who perform well on bad teams.
6. Finally, players have their positions reclassified from guards, forwards, and centers, to backcourt (guards) and frontcourt (forwards and centers).

This will allow us to factor in positions of players in a nontraditional way (guard, forward center). Based on the positionless nature of modern basketball, this change allows the versatility and diverse skills of frontcourt players to be emphasized slightly over traditional center skill such as paint presence and rebounding. Players with multiple listed positions on Basketball Reference were considered on a case by case basis.

At the conclusion of our data processing, we have a training data set of 1352 rows and 18 columns (16 used in the model), and a test data set of 343 rows and 16 columns.

4. Modeling

Two models were done in this project, linear regression and random forest regression. Linear regression is simple and straightforward, whereas for the random forest, hyperparameter tuning in the form of grid search was used to find the best model.

As stated before, the variables used (per 100 possessions) are: points, assists, rebounds, steals, blocks turnovers, field goal, three point field goals, and free throw, attempts, and percentages and team rank.

4.1. Linear Regression

The linear regression equation we use in our problem is:

$$\begin{aligned} Y = & \beta_0 + \beta_1 PTS + \beta_2 AST + \beta_3 OREB + \beta_4 DREB + \beta_5 STL + \beta_6 BLK + \beta_7 TOV + \beta_8 FGA + \beta_9 3PA + \beta_{10} FTA \\ & + \beta_{11} FG\% + \beta_{12} 3FG\% + \beta_{13} FT\% + \beta_{14} TEAMRANK + \beta_{15} POS + \beta_{16} POS \cdot PTS + \beta_{17} POS \cdot AST + \\ & \beta_{18} POS \cdot OREB + \beta_{19} POS \cdot DREB + \beta_{20} POS \cdot STL + \beta_{21} POS \cdot BLK + \beta_{22} POS \cdot TOV + \beta_{23} POS \cdot FGA + \\ & \beta_{24} POS \cdot 3PA + \beta_{25} POS \cdot FTA + \beta_{26} POS \cdot FG\% + \beta_{27} POS \cdot 3FG\% + \beta_{28} POS \cdot FT\% + \beta_{29} POS \cdot TEAMRANK \end{aligned}$$

with Y being All NBA vote share. Although daunting, this is simply a regression on vote share with all other variables as predictors, and an interaction term on POS (position). This means we have one model for frontcourt and one model for backcourt players. Intuitively it makes sense that certain stats will be more important for guards and others for forward/centers. Assists are viewed as more important for backcourt players while numbers like blocks and rebounds are more valued in frontcourt players.

Using this model, the predicted results for All-NBA Teams for the 2021-22 Season are:

Player (F/B)	Team (Rank)	Predicted Vote Share
First Team		
Nikola Jokic (F)	Denver Nuggets (11)	1.03
Joel Embiid (F)	Philadelphia 76ers (7)	1.00
Giannis Antetokounmpo (F)	Milwaukee Bucks (8)	0.99
Luka Doncic (B)	Dallas Mavericks (5)	0.86
Trae Young (B)	Atlanta Hawks (16)	0.61
Second Team		
Ja Morant (B)	Memphis Grizzlies (2)	0.55
Devin Booker (B)	Phoenix Suns (1)	0.51
LeBron James (F)	Los Angeles Lakers (23)	0.48
Kevin Durant (F)	Brooklyn Nets (14)	0.44
Jimmy Butler (F)	Miami Heat (3)	0.42
Third Team		
Stephen Curry (B)	Golden State Warriors (4)	0.47
Karl-Anthony Towns (F)	Minnesota Timberwolves (13)	0.42
Bam Adebayo (F)	Miami Heat (3)	0.42
Chris Paul (B)	Phoenix Suns (1)	0.41
Jayson Tatum (F)	Boston Celtics (6)	0.40

4.2. Random Forest Regression

For the random forest, we use the same predictors as in the linear regression. We perform hyperparameter tuning, with a 10-fold cross validation repeated twice, testing `mtry` values from 1 to 10 (number of variables randomly sampled as candidates at each split), using default `splitrule` based on variance and `min.node.size` of 5 (designates least amount of nodes in a leaf). In the end, we have that the optimal random forest has 1000 trees (this was fixed) with `mtry` = 10. Results from the hyperparameter tuning are displayed in this

plot:

figures/gridsearchresults.jpg

Figure 3: `mtry= 10` results in lowest RMSE

The model predictions for the 2021-22 All-NBA teams are as follows:

Player (F/B)	Team (Rank)	Predicted Vote Share
First Team		
Giannis Antetokounmpo (F)	Milwaukee Bucks (8)	0.80
Luka Doncic (B)	Dallas Mavericks (5)	0.74
Joel Embiid (F)	Philadelphia 76ers (7)	0.74
Nikola Jokic (F)	Denver Nuggets (11)	0.70
Ja Morant (B)	Memphis Grizzlies (2)	0.70
Second Team		
Kevin Durant (F)	Brooklyn Nets (14)	0.53
Jayson Tatum (F)	Boston Celtics (6)	0.48
Devin Booker (B)	Phoenix Suns (1)	0.45
Donovan Mitchell (B)	Utah Jazz (9)	0.43
LeBron James (F)	Los Angeles Lakers (23)	0.39

Third Team		
Stephen Curry (B)	Golden State Warriors (4)	0.39
DeMar DeRozan (F)	Chicago Bulls (12)	0.36
Trae Young (B)	Atlanta Hawks (16)	0.34
Jimmy Butler (F)	Miami Heat (3)	0.33
Rudy Gobert (F)	Utah Jazz (9)	0.33

4.3. Model Analysis and Comparison

Looking at the predictions from both models, they 4 of the same First Team members, and 8/15 players in the same team spots. Twelve of the fifteen total players appear on both (disregarding teams) with the linear regression slotting Bam Adebayo, Karl-Anthony Towns, and Chris Paul into the Third Team while the random forest has Rudy Gobert and DeMar DeRozan in the Third Team with Donovan Mitchell as Second Team (surprising apperance!). These results suggest that our models are pretty similar, and predict pretty well the First and Second Teams; usually these are more clear cut, with a variety of players deserving consideration for Third Team. What really is important here, (and what affirms the NBA’s decision to make Embiid and Jokic eligible as forwards) is that both models seem to agree that Jokic, Giannis, and Embiid are the three best frontcourt players, all deserving of first team spots, ignoring positional conflicts at center.

One point of interest is that the random forest is able to keep vote share between 0 and 1; whereas linear regression’s pitfall ends up being that it can predict negative values for players, something that is not possible in real life (we can’t have negative votes). It also predicts values over 1 six times, 08-09 and 09-10 LeBron, 15-16 Curry (he was also unanimous MVP), 18-19 and 19-20 Giannis, and 19-20 Harden, again impossible (also no Michael Jordan?).

Besides comparing the model predictions, we can also look at some characteristics of the model, namely adjusted- R^2 and Root Mean Squared Error (RMSE). The linear regression has an adjusted- R^2 0.5803, which means that 58.03% of the variation in our training data’s player vote share can be explained by the variation in our predictors. For the random forest,

the R^2 is 0.6362 indicating that 63.62% of the variation in vote share is due to variation in our predictors. For the linear regression and random forest model, the RMSE is approximately 0.197, and 0.189, respectively. RMSE measures the average distance of the predicted vote share to the actual vote share in the training set; a smaller number implies less variation and thus more “confidence” in our predictions for the training data.

Looking at both R^2 and RMSE suggests that the random forest model is slightly better, but the problem with these metrics is that they are not done on test data, but rather the data we train on (in the random forest case it is on out of bag data in cross validation). In fact, we really have no valid test data, since we don’t have the results of the All NBA teams for the 2021-22 season. Instead, another way we can analyze and compare our models is to check out the coefficients and see how they are being used — if they are being used in ways that match our supposed understanding of the problem.

For linear regression, we can look at the coefficients of our predictors, as well as their significance in the model. Besides our predictors that are factors (team rank and position), the predictors of at a 0.01 significance level are defensive rebounds, assists, steals, turnovers, and points, with all of them being positively associated with vote share besides turnovers. Notably, the coefficient for points is approximately 0.31, meaning that all else equal, a player whose scoring increases an entire standard deviation compared to the rest of the league (that has played at least half the games) will see his All-NBA vote share increase on average by 0.31. In the 2021-22 NBA Season, one standard deviation was 6.16 points per 100 possessions, with the mean being 21.38; Joel Embiid led the league in 45.1 points per 100 possessions. First, this is a huge number, and it also makes a lot of sense; scoring is the name of the game, and regular season awards will always lean towards players with high scoring output. It’s interesting that none of the shooting attempts or percentages ended up being very important, with field goal and free throw percentage even having a negative coefficient. This is probably because the best players end up having to take more difficult shots compared to sharpshooters due to defensive gameplanning, and this data can be very noisy across seasons, making it hard for us to see the effects of the “better efficiency = better player” philosophy. As long as you don’t shoot terribly (think low 40% or worse), you should be fine as a player. Next, looking at the coefficients for interactions with frontcourt position, assists, and steals have

their coefficients decrease slightly, with blocks, offensive rebounds, and points increasing as well as turnovers (meaning the negative impact of turnovers is smaller). This is consistent with traditional perceptions of what players should be good at; for guards it's more important to be good at passing and ball security, while frontcourt players are expected to rebound better, protect the paint, and to be a little more loose with the ball. Finally, we can look at team rank, which has only one positive coefficient, `rank=3`; but more importantly, as rank decreases, the coefficients continue to increase, reaching consistently over -0.12 after the team rank is out of the top 10, and peaking at -0.33 (`rank=28`). Interpreting this, if two players had identical stats but one of them was on the best team and one of them was on the worst team, the player on the better team would have on average a higher vote share by 0.33. Overall, the linear regression model assigns positive weights to predictors that are generally associated with individual player success, as well as factoring in slight changes for position and team rank.

Now for random forest, we can use gini impurity feature importance. This measures the total decrease in node impurity averaged over all trees in the forest. Node impurity tells us the probability of misclassifying an observation, so the larger the decrease, the lower chance we have of misclassifying. Thus the higher the gini impurity feature importance, the better and more important it is in arriving at the correct classification in our model. Using `ranger_importance`, we have that Points, Rank (Team performance), Assists, and Rebound are most important, with feature importances of 42, 19, 13, 10, and 8, respectively. Once again, scoring has been the most important historic predictor in winning All-NBA. At the same time, the best scorers in the league excel at drawing contact and getting sent to the free throw line. Assists and Rebounds are the other part of traditional stat lines, so they are important to the model as well. We also see rank as the second most important feature, but upon further examination, this could just be a case of correlation rather than causation (or the chicken and egg problem). Players are not selected as All-NBA because their team is good like the model importance suggests, rather, because they are All-NBA players, they are responsible for the success of their team; here the model could simply be reflecting this idea. One final note is that it is interesting to see Position ranked as the least important feature; intuitively this might seem very important as different stats are valued for different

positions. Like the linear regression model, the random forest model recognizes predictors that are traditionally associated with individual player success, but two pitfalls are that we cannot easily interpret the impact of individual predictors due to the complexity of the model and feature importance is a reflection of trends in our historical training data, not necessarily on how modern voters actually select All-NBA players.

Finally, let's take a look at some controversial All-NBA decisions that arise with the debate of positions, and see what our models predict. In the 2015-16 season, with weak performances from center, DeAndre Jordan earned First Team All-NBA, with many believing Draymond Green, a small ball 5, to be snubbed for First Team. Our model results are shown below in comparison to the real results:

First Team (Actual, LM Prediction, RF Prediction)					
Stephen Curry (B)	1	Stephen Curry (B)	1.06	Stephen Curry (B)	0.91
LeBron James (F)	0.99	LeBron James (F)	0.77	LeBron James (F)	0.94
Russell Westbrook (B)	0.97	Russell Westbrook (B)	0.77	Russell Westbrook (B)	0.79
Kawhi Leonard (F)	0.89	Kawhi Leonard (F)	0.53	Kawhi Leonard (F)	0.73
DeAndre Jordan (F)	0.49	Kevin Durant (F)	0.78	Kevin Durant (F)	0.75
Second Team (Actual, LM Prediction, RF Prediction)					
Kevin Durant (F)	0.7	Paul George (F)	0.36	DeAndre Jordan	0.43
Draymond Green (F)	0.67	Draymond Green (F)	0.41	Draymond Green (F)	0.44
DeMarcus Cousins	0.43	DeMarcus Cousins	0.46	DeMarcus Cousins	0.32
Chris Paul (B)	0.55	Chris Paul (B)	0.69	Chris Paul (B)	0.57
Damian Lillard (B)	0.34	James Harden (B)	0.44	Damian Lillard (B)	0.37

Third Team (Actual, LM Prediction, RF Prediction)					
Andre Drummond (F)	0.27	Andre Drummond (F)	0.33	Andre Drummond (F)	0.28
Paul George (F)	0.24	DeAndre Jordan (F)	0.3	Paul George (F)	0.26
LaMarcus Aldridge (F)	0.16	Hassan Whiteside (F)	0.33	LaMarcus Aldridge (F)	0.16
Kyle Lowry (B)	0.24	Kyle Lowry (B)	0.37	Kyle Lowry (B)	0.26
Klay Thompson (B)	0.25	Isaiah Thomas (B)	0.33	Klay Thompson (B)	0.21

Looking at the results, the random forest model actually gets everything same with the actual All-NBA squad, except Kevin Durant and DeAndre Jordan are swapped in position; was Durant the real player snubbed compared to the Draymond? He did have the next highest vote share among frontcourt players. Meanwhile, the linear regression gets a little crazy, inserting James Harden, Hassan Whiteside and Isaiah Thomas into the All-NBA teams, probably due to their high counting stats. But both models (and technically the actual results) are similar in that they agree DeAndre Jordan is not All-NBA First Team worthy.

Another controversial All-NBA choice as previously mentioned, was during the 2019-20 season, when Khriston Middleton was left off the All-NBA Third Team for Ben Simmons despite receiving more votes. Both models agreed with this sentiment, placing Middleton on the Third Team, and having Ben Simmons off.

Lastly we will take a look at the frontcourt results for the 1993-94 season. In this season, Patrick Ewing missed the All-NBA team behind centers Hakeem Olajuwon, David Robinson, and Shaquille O'Neal, despite the fact that Ewing was 5th in MVP voting. The idea that the fifth best player in a season can miss All-NBA again raises questions regarding the All-NBA Teams expanding past positional limitations. This table shows the frontcourt results:

First Team (Actual, LM Prediction, RF Prediction)					
Hakeem Olajuwon	0.65	Hakeem Olajuwon	0.84	Hakeem Olajuwon	0.78
Karl Malone	0.81	David Robinson	0.94	David Robinson	0.75
Scottie Pippen	0.94	Shaquille O'Neal	0.79	Scottie Pippen	0.75

Second Team (Actual, LM Prediction, RF Prediction)					
David Robinson	0.66	Patrick Ewing	0.53	Karl Malone	0.66
Charles Barkley	0.4	Charles Barkley	0.56	Charles Barkley	0.43
Shawn Kemp	0.41	Shawn Kemp	0.5	Shawn Kemp	0.38
Third Team (Actual, LM Prediction, RF Prediction)					
Derrick Coleman	0.27	Scottie Pippen	0.49	Patrick Ewing	0.35
Dominique Wilkins	0.21	Karl Malone	0.43	Dominique Wilkins	0.22
Shaquille O'Neal	0.18	Chris Webber	0.32	Shaquille O' Neal	0.37

Looking at our two models, they both have Patrick Ewing as an All-NBA mention, with the linear regression having Ewing in the Second Team and the random forest with him in the Third. The linear regression really throws positions out the way, with the four centers (Hakeem, Shaq, Ewing, Robinson) in the four out of the top five frontcourt spots in terms of vote share. The random forest is a little more conventional, with 8 out of the 9 frontcourt spots being the same, only swapping out Derrick Coleman for Patrick Ewing.

Overall, both models do a good job predicting All-NBA locks, players that clearly were a step above their peers during the regular season. However, when it comes to more fringe candidates and actual ordering of team placements, the models have clear tendencies. The linear regression gives more weight to individual counting statistics (points, rebounds, assists) as you can see with the three best centers all First Team in 93-94, or with predicting Trae Young as First Team this year. Meanwhile the random forest seems to have a more holistic view, closer to actual All-NBA results with more weight towards defense, team rank, and with a little more stricter positional focuses.

5. Historical Changes

A fun thought exercise is if we used the models to predict all the past All-NBA teams — what would change in terms of career accolades for these players? Here we the top 50 players (in terms of All-NBA selections from 88-89 onwards) and changes to their All-NBA nods based on the models.

Player	Real	LM	RF	Player	Real	LM	RF
LeBron James	17	17	15	Damian Lillard	6	4	5
Kobe Bryant	15	13	14	Jason Kidd	6	5	7
Tim Duncan	15	18	15	Patrick Ewing	6	8	9
Shaquille O' Neal	14	15	16	Paul George	6	4	3
Karl Malone	13	15	15	Amar'e Stoudemire	5	4	5
Dirk Nowitzki	12	12	12	Ben Wallace	5	0	2
Chris Paul	10	12	9	Blake Griffin	5	4	6
David Robinson	10	12	10	Chris Webber	5	4	5
John Stockton	10	13	13	Giannis Antetokounmpo	5	5	5
Gary Payton	9	8	7	Grant Hill	5	4	3
Hakeem Olajuwon	9	10	10	Kawhi Leonard	5	6	6
Kevin Durant	9	9	9	Kevin Johnson	5	6	4
Kevin Garnett	9	9	8	LaMarcus Aldridge	5	0	4
Russell Westbrook	9	10	10	Mitch Richmond	5	0	0
Charles Barkley	8	9	8	Tim Hardaway	5	4	3
Dwight Howard	8	7	8	Yao Ming	5	4	6
Dwyane Wade	8	9	9	Anthony Davis	4	5	6
Michael Jordan	8	9	9	Chris Mullin	4	1	2
Allen Iverson	7	7	7	Clyde Drexler	4	6	6
James Harden	7	8	8	Dominique Wilkins	4	2	5
Scottie Pippen	7	7	6	Jimmy Butler	4	3	2
Stephen Curry	7	8	7	Mark Price	4	5	4
Steve Nash	7	6	8	Pau Gasol	4	4	2
Tracy McGrady	7	6	6	Paul Pierce	4	8	3
Carmelo Anthony	6	6	8	Rudy Gobert	4	2	1

Looking at the top players, both models have similar numbers for All-NBA selection, with Tim Duncan taking the top spot in the linear model and Shaq in the random forest. From the models, we can see that dominant big men of the 90s and early 2000s eras see more All-NBA nods. Hakeem, Shaq, Robinson, and Ewing all receive some extra nods, as under the models, they are competing with other frontcourt players rather than each other at center. Not in the graph is also Alonzo Mourning, who gets 5 selections in both models compared to 2 in real life; Shawn Kemp also goes from 3 to 6 and 5 in the linear regression

and random forest model, respectively. On the opposite end, centers during a weaker time of big men see less All-NBA selections. Ben Wallace, defensive savant on the Pistons, goes from 5 selections to 0 in the regression model and 2 in the random forest, and Dikembe Mutombo goes from 3 to 0 both. Rudy Gobert decreases as well (he is last in the table). Some players taking their spots are Chris Bosh, who has 4 selections under the linear regression (he has 1 in real life), and Carlos Boozer, who has 4 under the random forest (he also has just 1). Arvydas Sabonis, a talented passing big man who joined the league after his prime playing days in Europe, has 3 selections in the linear regression, compared to his actual 0. Other more defensive oriented centers that lose their selections are Andrew Bogut and Al Horford.

Since the models use per possessions statistics, sixth men, who do their scoring in shorter bursts and less time, are valued more highly in the models. Manu Ginobili has 5 and 3 selections in the linear regression and random forest, respectively, compared to 2 in real life. Lou Williams, who won Sixth Man of The Year averaging 20 points in 26 minutes for the 18-19 Clippers, gets an All-NBA nod as well in the random forest. This insight would be useful in the volume-efficiency discourse; that because a player plays less minutes, they can go maximum effort and be more efficient in that limited amount of playing time. Could a player sustain his rate of scoring if he played more?

Two more notable jumps: Paul Pierce in the linear regression has 8 All-NBA selections compared to 4 in real life, and 3 in the random forest. A prolific small forward whose prime coincided with other great forwards (LeBron, Melo, KD, Dirk, KG, Duncan, etc...), he sees more nods under the linear regression as the model deems him to have outperformed the centers of his era based on his scoring and success with the Celtics. Stuck in the basketball purgatory of Sacramento in his early career, DeMarcus Cousins is another prime beneficiary of the models; the linear regression and random forest have him on an All-NBA team 4 and 6 times respectively, compared to his actual 2. This discrepancy is partially the “empty calories” argument of a player putting up big numbers on a bad team, and also the perception that DeMarcus Cousins was a hot head and “locker room cancer”, which negatively impacted voter perception. The models however, do not see this, and think his numbers speak enough for his level of play.

Overall, from historical changes based on the models, we can see that both models tend

to value offense over defense relative to past voter behavior. The linear model can go a bit to the extreme with positional flexibility, really pushing the limits of the All-NBA Team being the best 15 players and greatly valuing scoring and other box score numbers. The random forest model is a little more conventional, agreeing more often with voting history, but still reflecting league trends of positional strength in its results. In summary, historic voting trends stuck strictly to three centers, whereas the linear model really emphasizes volume stats, scoring, and maximum positional flexibility with the top 15 players in the season, and the random forest is a happy medium between the two.

6. Final Thoughts

Voting for All-NBA is not easy. It's a tough task selecting the 15 best players in a season especially when the talent level is at an all time high. It's also a uniquely human task, with each voter having different criteria and biases that dictate their decisions (as we have seen in several examples mentioned). At the same time, it's an equally difficult task to try to predict All-NBA, especially with the constraints of machine learning. The model learns on past data and is also partially evaluated on how well it performs on this past data. This begs the question — how much of our model is just reflections of past data and how much of it is a unique and fresh take on All-NBA voting?

Ultimately, the models in this project are just a means to explore a thought experiment — what if we removed a positional limitation so that All-NBA could better highlight the diverse skills of players in the modern game? Just like there is no “right” or “correct” way to vote for All-NBA, there is no “best” model, in terms of traditional metrics like R-squared or accuracy. We *want* the models to be different from past voting records, but how much until the model results are *too* different? It's a unique challenge because there is no right answer to whether we should switch our voting to a two backcourt/three frontcourt system. But hopefully, these models and their results brought some insight to the All-NBA voting system, and how we might proceed as the league continues to expand and grow.