NBA Awards Prediction - Part 2

Ryan Drost
Luigi Noto

Methodology

We explored a few areas of additional analysis. First, we looked to improve our logistic regression model for year-by-year All-NBA award prediction. We added data from the 2022 season and used this along with 2021 as the test set. We particularly wanted to use two seasons in our test set since there is likely a good amount of variance in the model's predictive ability, depending on which players earn All-NBA awards in a given season; some seasons are inherently much easier to predict than others. We also investigated the calibration of the model by determining the sum of the model's predicted probabilities for all players in each season.

Moreover, we trained a random forest for year-by-year All-NBA award prediction using the same features engineered for the logistic regression model, after performing a cross-validation over a grid of hyperparameter values, in order to see whether performance can be improved using a nonlinear model and to investigate the relative importances of the engineered features in the prediction by means of impurity-based feature importance.

Additionally, we created a linear regression model to predict future All-NBA awards a player will win over the rest of their career. We approached this in a similar manner to our year-by-year predictions, using data available prior to the start of each season and engineering features to predict these results. We only used player seasons from 1990 through 2005 for this model. The later data is problematic because it includes current players, which means we do not have a ground truth value for their future All-NBA appearances on which to train the model.[1] In seasons prior to 1990, only two All-NBA teams were chosen (as opposed to three every year since) which would affect our model negatively for current predictions. We split these datapoints randomly into train and test sets to build the model.

Results

For our final logistic regression model for year-by-year All-NBA prediction, we have 10 features including our baseline binary predictor variable of whether or not the player made the All-NBA team the season before (Prev_1_AllNBA). The metrics for the model performance are updated here.[2] Notably, a calibration score of 15 is ideal since we know that there are 15 awards each season, and our final model is very good in this respect as the sum of the individual probabilities for the 2021 and 2022 season are 15.41 and 14.93, respectively (see Figure 1).

The random forest model for year-by-year All-NBA prediction achieved very similar performance metrics as the final logistic regression model. Figure 2 shows the impurity-based importances of

[1] There are a couple active players included, but most have a very negligible chance of making any more All-NBA teams. LeBron James is the obvious exception, but we decided it was worth using several more years of training data at the expense of one player having a possibly inaccurate ground truth number.

[2] The main reason for the changes is the increased size of the test set, but we did also correct some data inconsistencies and did some additional feature engineering.

the 10 predictor variables. We can see that the baseline predictor variable (Prev_1_AllNBA) and a closely related variable (Prev_1_AllNBAV) have the largest importance, but also other variables have a substantial influence on the prediction, in particular Prev_1_WS_g, which represents the player's win shares per game in the previous season.

Metric	Baseline LR Model	Final LR Model	Random Forest Model
AUROC	0.760	0.982	0.974
Accuracy	0.976	0.977	0.981
F1 Score	0.533	0.571	0.593
Precision	0.533	0.737	0.667
Recall	0.533	0.467	0.533
Log Loss	0.084	0.056	-
Calibration	16.84	15.17	15.18

Our final model for predicting future All-NBA appearances includes 11 features and has $R^2 = 0.28$ on the test set. This is not particularly high, but for a challenging problem it presents reasonable predictions that make sense intuitively. Our baseline model which included two variables (age and number of All-NBA teams in the past 3 seasons) had $R^2 = 0.15$ so this final model represents a significant improvement.

Analysis

As expected, adding a season to the test set affected our model performance and demonstrates the variance in the difficulty of predictions from year to year. In particular, improving recall still proved very challenging. It is simply quite difficult for our logistic regression model to identify players who will make an All-NBA team when they have not already made one, without predicting many more incorrect positives. Some of these false negatives are just below the 0.5 probability threshold, but any player who has never made an All-NBA team has a very low probability. We also did not end up adding any features to differentiate predictions among rookies. This was strategic because only four rookies have ever made an All-NBA team, so the very low predicted probability is realistic; while slightly unsatisfying, we determined ultimately that it is only a minor issue.

Our linear model produces intuitive results but also suffers from a similar issue with hard cut-offs. Although we have engineered features to correct for this as best as possible, ultimately many of them are binary variables and if a player does not score a positive in any of them, they will have a very low prediction. We do account for rookies in this model with a separate linear regression model to predict future All-NBA appearances by rookies given their draft pick number, which we then use as a feature for this model. This proved more necessary for this model as it helped the model significantly. This makes sense because rookies, particularly highly drafted ones, are fairly likely to make one or more All-NBA teams in their career, even if they are highly unlikely to make it in their rookie season.

Because we used a linear regression, some players have a negative prediction which is of course not possible. While unsatisfying, the lowest prediction is -0.8 (and the second lowest is -0.3), and we can simply interpret a negative prediction as zero. We experimented with using a Poisson regression to avoid this issue, but it produced much more skewed results which did not make nearly as much sense intuitively. The calibration of the model also suffers because the training data is from 20 years ago when there were between 300 and 350 players in the league each season. In 2023, there have been over 500 players which causes the sum of each season's predictions to be much higher. Despite these shortcomings, the model tackles a challenging question and produces reasonable results. The biggest flaw is that it still struggles to identify star players before they are stars which would be the most useful aspect, since it is also very difficult for people to predict that in general.

Appendix

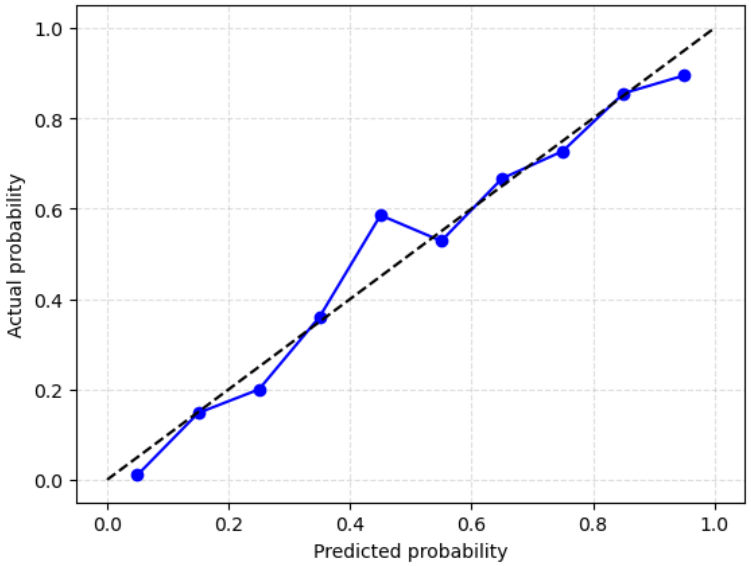

Figure 1: Calibration of training set for logistic model

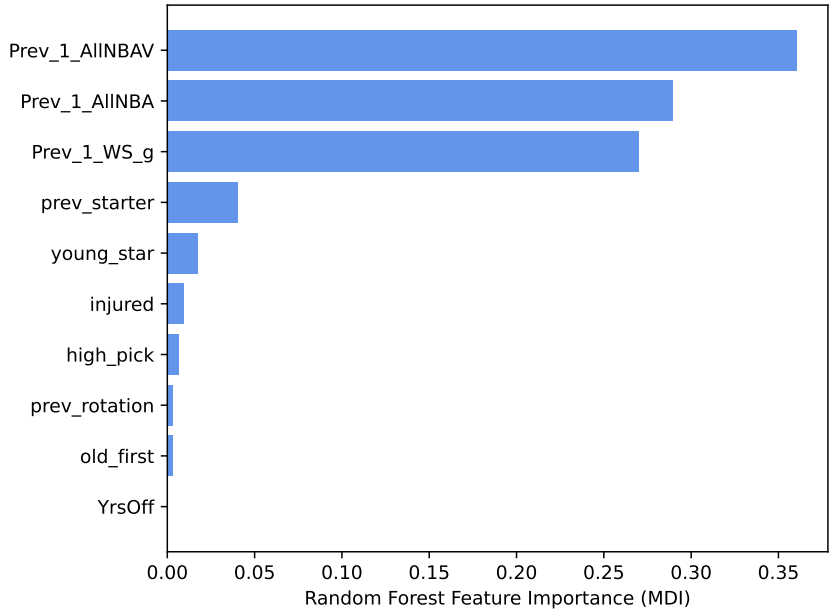

Figure 2: Impurity-based importances (mean decrease in impurity) of random forest model for year-by-year All-NBA prediction

Logistic Model (Year-by-Year) Features

Feature	Type	Coefficient	Explanation
Prev_1_AllNBA	Binary	2.05	Made All-NBA Team in previous season
Prev_1_AllNBAV	Discrete	0.22	Value metric for All-NBA team from previous season: 10 for first team, 5 for second team, 1 for third team, 0 for none
old_first	Binary	-0.36	Age >= 28, Made All-NBA team for the first time in the previous season, but was not on first team
young_star	Binary	0.52	Age < 25, Made All-NBA first or second team in any season previously
prev_rotation	Binary	-1.21	Played < 15 minutes per game in previous season
prev_starter	Binary	-2.72	Played < 28 minutes per game in previous season
YrsOff	Continuous	-1.69	Number of seasons missed since the player's most recent season
Prev_1_WS_g	Continuous	5.76	Win shares per game in previous season
injured	Binary	1.14	Fewer than 50 games in previous season with more than 15 points per game and Prev_AllNBAV_decay > 10 (a metric based on AllNBAV with beta decay)
high_pick	Binary	2.15	Drafted in the top 10, made All-Rookie first team, and fewer than 3 years of experience

Linear Model (Future All-NBA Awards) Features

Feature	Type	Coefficient	Explanation
Age	Discrete	-0.04	Player's age
Prev_3_AllNBAV	Discrete	0.07	Value metric for All-NBA team from previous 3 seasons: 10 for first team, 5 for second team, 1 for third team, 0 for none
young_star	Binary	1.45	Age < 25, Made All-NBA first or second team in any season previously
v_young_star	Binary	6.08	Age < 23, Made All-NBA first or second team in any season previously
young_AS	Binary	4.64	Age < 25, Made at least 2 All-Star teams
no_play	Binary	-0.13	Played < 15 minutes per game in previous 3 seasons combined and in most recent season
Prev_3_AS	Discrete	0.15	Number of All-Star selections in previous 3 seasons
Prev_3_VORP_g	Continuous	10.09	VORP per game in previous 3 seasons
high_pick	Binary	1.07	Drafted in the top 10, made All-Rookie first team, and fewer than 3 years of experience
star	Binary	1.67	Drafted in the top 10, made All-Rookie first team, and fewer than 3 years of experience
rookie	Continuous	0.81	Metric for rookies (Exp = 0) based on separate linear regression of draft pick number and career All-NBA awards

2023 All-NBA Probabilities[3]

Player	Actual	Prediction	Probability
Jayson Tatum	Almost certain	Yes	0.879
Luka Doncic	Almost certain	Yes	0.873
Nikola Jokic	Almost certain	Yes	0.872
Giannis Antetokounmpo	Almost certain	Yes	0.863
Devin Booker	Unlikely (injury)	Yes	0.799
Ja Morant	Maybe	Yes	0.696
Joel Embiid	Almost certain	Yes	0.656
Kevin Durant	Maybe	Yes	0.624
Stephen Curry	Almost certain	Yes	0.586
DeMar DeRozan	Unlikely	Yes	0.573
Kawhi Leonard	Unlikely (injury)	Yes	0.506
Chris Paul	Unlikely	No	0.396
Karl-Anthony Towns	No (injury)	No	0.388
LeBron James	Almost certain	No	0.381
Trae Young	Unlikely	No	0.378

2023 Future All-NBA Award Predictions[4]

Player	Age	Prediction
Luka Doncic	23	11.55
Jayson Tatum	24	9.80
Trae Young	24	5.77
Giannis Antetokounmpo	28	5.43
Nikola Jokic	27	5.17
Joel Embiid	28	3.66
Anthony Davis	29	3.40
Kawhi Leonard	31	2.69
Ja Morant	23	2.53
LeBron James	38	2.18
Paolo Banchero	20	2.16
Zion Williamson	22	2.05
Stephen Curry	34	2.01
LaMelo Ball	21	1.91
Kevin Durant	34	1.75

[3]As is true throughout, we only used data available prior to the season. Although we don't know the actual All-NBA selections just yet, since the regular season is over we do have a much better idea of whether the predictions will be correct and that is noted in the table.

[4]Once again, we only used data available prior to the current season. Therefore these predictions include the 2023 season as part of the future. As soon as this season's All-NBA teams are announced, we could add that data and do true future predictions.

Highest All-NBA Probabilities

Players with probability ≥ 0.9 to make All-NBA in training (1989-2020) and test set (2021-2022)

Player	Season	Actual	Probability
Karl Malone	2000	Yes	0.982
Tim Duncan	2000	Yes	0.968
Tim Duncan	1999	Yes	0.964
Luka Doncic	2021	Yes	0.962
Anfernee Hardaway	1996	Yes	0.961
David Robinson	1992	Yes	0.956
Shaquille O'Neal	2000	Yes	0.945
Chis Paul	2009	Yes	0.927
Tracy McGrady	2004	Yes	0.924
LeBron James	2022	Yes	0.924
LeBron James	2007	Yes	0.920
Anthony Davis	2016	No	0.920
Tim Hardaway	2000	No	0.919
LeBron James	2009	Yes	0.919
Kevin Durant	2011	Yes	0.916
LeBron James	2006	Yes	0.911
Kevin Durant	2013	Yes	0.911
Tim Duncan	2001	Yes	0.906
Dwight Howard	2010	Yes	0.906
Michael Jordan	1989	Yes	0.902
LeBron James	2014	Yes	0.900

Highest Future All-NBA Award Predictions

Players predicted to have ≥ 5.5 future All-NBA awards in the training set (1990-2005)

Player	Season	Actual	Prediction
Kobe Bryant	2001	13	13.63
Tracy McGrady	2004	4	11.28
Tim Duncan	2001	12	11.19
Kobe Bryant	2003	11	10.58
Tim Duncan	1999	14	10.44
Kevin Garnett	2001	7	9.81
Tracy McGrady	2002	6	8.93
Grant Hill	1997	4	8.59
Shaquille O'Neal	1997	11	8.01
Shaquille O'Neal	1996	12	7.96
Shaquille O'Neal	1995	13	7.08
Tim Duncan	2000	13	6.72
Michael Jordan	1990	7	5.89
Michael Jordan	1991	6	5.85
Kevin Garnett	2000	8	5.82
Michael Jordan	1992	5	5.74
Kevin Garnett	1999	9	5.64
Michael Jordan	1993	4	5.62