Modeling All-NBA Team Voting

Eric J. Wissner

gtID: ******824

ejwissner@gatech.edu

April 16, 2023

ISYE 7406 - Data Mining and Statistical Learning

Abstract

Despite advances in how analytics are used to understand and measure National Basketball Association player performance, "All-NBA Team" honors are determined by voting conducted by a panel of 100 sports reporters and broadcasters. In this project, historical player statistics have been combined with annual voting results to develop an analytic model representing the factors those voters have used over the years, consciously or subconsciously, to identify the league's top players. That model was then used to assess the most recent selections for the 2021-2022 season, based only on player data for that season.

The data used in this analysis was sourced from Sport Reference and includes twenty years of player statistics, advanced/composite measures, and annual All-NBA Team selections. Data for the first nineteen years was used to train and validate several models and the last season's worth was held for the assessment (prediction) noted above. After removing records for players with fewer than 10 games played per year, there were 8,223 records in the training/validation data set.

Analytic models evaluated included: logistic regression, LASSO, K-nearest neighbors (KNN), KNN after principal component analysis, classification tree, random forest, and boosting. Techniques were employed to tune parameters and 10-fold cross validation was used to evaluate model performance. The boosted model was identified as the top performer.

Using player statistics from the 2021-2022 NBA season, the final model agreed with 13 of the 15 players that were named to the All-NBA Team while suggesting that Rudy Gobert and Jimmy Butler should have been included in place of Pascal Siakam and Chris Paul.

Introduction

Statistics have long been associated with sports. Prior to the internet and ESPN, newspapers reported box scores for games played the previous day. Record books continue to memorialize great players and their amazing achievements: Joe DiMaggio's 56 game hitting streak, Ted Williams's 0.406 batting average, and Wilt Chamberlain's 100-point game are just a few of the sports records that are highly revered by fans (Fox Sports, 2016).

In recent years, analytics have taken a more prominent role across the different sports leagues. Moneyball (the book in 2003 and the movie in 2011) introduced the general public to the concept of teams beginning to leverage analytics to optimize personnel decisions (Doll, 2022). The National Basketball Association has also embraced analytics and almost every team now has an analytics department. The league collects data on player- and ball-movement 25 times per second and teams use data to make decisions on player matchups and when to rest players. Data analytics has also been cited as a reason why teams are taking more three-point attempts than in previous years (Merrimack College, 2021).

At least one annual decision, however, is still made in the same way it was back in 1946. For every year the NBA has been around, a panel of sports writers and reporters have voted to determine who will be named to the All-NBA Teams (Vorkunov, 2022). Three teams of five players are named, with the highest vote recipients for each position earning "first" team honors.

Problem Statement

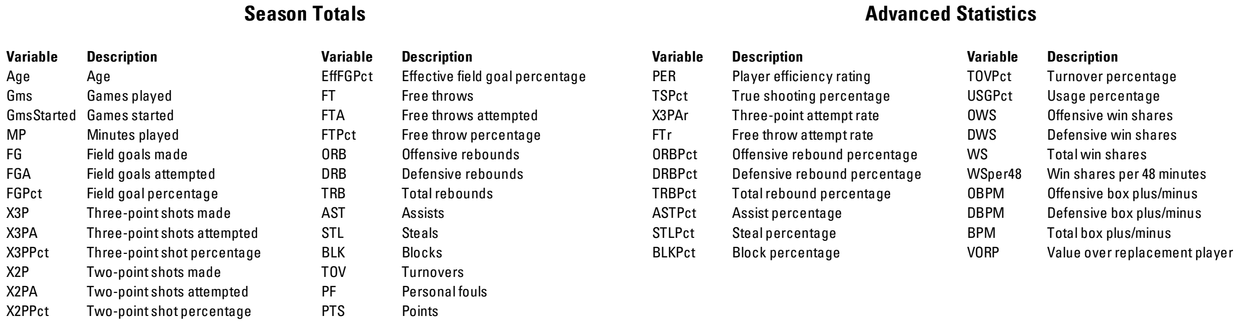
This analysis sought to determine if historical All-NBA Team voting results could be modeled using individual player statistics for each respective season. Such a model would represent the factors that the voters collectively considered, consciously or subconsciously, when identifying each season's top players.

That model could then be used to assess the All-NBA Team voting results for the 2021-2022 season. Which players were voted on to the team consistent with the patterns used by previous voters? Which players that were not selected should have received the end-of-year honors based on the data?

Data Source

Data for this analysis was obtained from Sports Reference's basketball-reference.com website (Sports Reference, n.d.) and included three types of information: basic totals per player per season, advanced metrics and composite scores per player per season, and the results of the annual All-NBA Team voting. Web scraping and regular expressions were employed in a Python script to construct a data set that included 20 years of player performance and award data.

The data set included the following predictor variables:



Season Totals

Variable	Description	Variable	Description
Age	Age	EffFGPct	Effective field goal percentage
Gms	Games played	FT	Free throws
GmsStarted	Games started	FTA	Free throws attempted
MP	Minutes played	FTPct	Free throw percentage
FG	Field goals made	ORB	Offensive rebounds
FGA	Field goals attempted	DRB	Defensive rebounds
FGPct	Field goal percentage	TRB	Total rebounds
X3P	Three-point shots made	AST	Assists
X3PA	Three-point shots attempted	STL	Steals
X3PPct	Three-point shot percentage	BLK	Blocks
X2P	Two-point shots made	TOV	Turnovers
X2PA	Two-point shots attempted	PF	Personal fouls
X2PPct	Two-point shot percentage	PTS	Points

Advanced Statistics

Variable	Description	Variable	Description
PER	Player efficiency rating	TOVPct	Turnover percentage
TSPct	True shooting percentage	USGPct	Usage percentage
X3PAr	Three-point attempt rate	OWS	Offensive win shares
FTr	Free throw attempt rate	DWS	Defensive win shares
ORBPct	Offensive rebound percentage	WS	Total win shares
DRBPct	Defensive rebound percentage	WSper48	Win shares per 48 minutes
TRBPct	Total rebound percentage	OBPM	Offensive box plus/minus
ASTPct	Assist percentage	DBPM	Defensive box plus/minus
STLPct	Steal percentage	BPM	Total box plus/minus
BLKPct	Block percentage	VORP	Value over replacement player

A binary indicator of whether each player earned All-NBA Team honors in that year served as the response variable.

Exploratory Data Analysis

Prior to training the classification models, exploratory data analysis was conducted to better understand the data and to identify outliers or other anomalies. One issue involved records for players who only played in a handful of games in a season.

Although there is no "minimal games played" requirement for All-NBA Team consideration (Helin, 2023), data for players with less than 10 games played per season were removed due to the extreme values those records included. For instance, values for some basic statistics were often zero for these data points while percentage-based variables were sometimes close to 100% - technically accurate but misleading.

Methodology

After removing records for players with less than 10 games played per season, the total data set included 8,717 records. 8,223 of these records were for the first nineteen years of data and were allocated for use as the training/validation data set, leaving the remaining 494 records for the 2021-2022 season as the prediction data set.

When training the data, different techniques were employed *in conjunction with 10-fold cross-validation* to tune parameters associated with each of the individual models. The following provides a summary of each analytic model as well as the tuning methods applied:

> **Logistic Regression**. This is a machine learning model that returns the log odds (and subsequently, the probability) of a binary outcome. That probability can be interpreted as either of the outcomes based on a threshold value. Typically, that value is 0.50. Any probability returned above that number becomes a "1" (or an

All-NBA Team member in our analysis) and anything below becomes a “0” (no award).

LASSO. This regularization technique minimizes the values of model coefficients in a manner that can reduce some of them to zero, effectively removing them from the model. The lambda parameter (λ) was tuned to return a mean squared error within 1 standard error of the smallest cross-validation error. This is not necessarily the smallest CV error; but tends to return variables that later perform well on unseen data.

K Nearest Neighbors (KNN). In this method, predictions are made based on the data points (quantity of “k”) nearest to the data points for the record in question. Those neighbors effectively vote and the most frequently occurring classification becomes the prediction for that record. The model can be tuned by using different values for k (in other words, considering a different number of neighboring data points). In this analysis, odd values between 3 and 15, inclusive, were evaluated.

Principal Component Analysis with KNN. In this technique, a data set’s original variables get replaced by a set of principal components (PCs), ranked in descending order based on total variance. The KNN method is then used to train a model based on the response variables and these surrogate variables. This model was tuned to include the first PCs that cumulatively represented at least 80% of the variance in the data and to identify the optimal value for k as in the KNN section above.

Classification Tree. This model uses a treelike approach to determine classifications through successive segmentation of values for different variables within the data set.

Random Forest. This ensemble approach builds on the traditional classification tree model by creating a large number of trees that include different (repeated) rows from the original training set as well as random subsets of all predictors. For this analysis, both the number of created trees and the number of predictor variables to include were tuned.

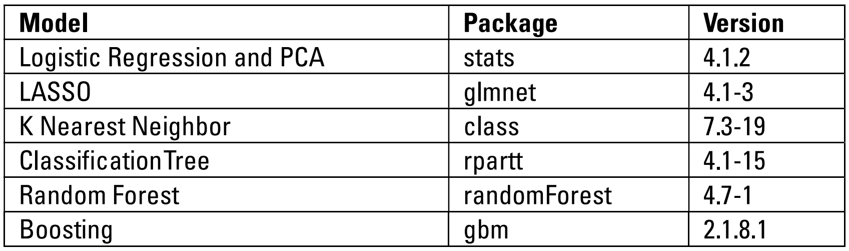
Boosting. Boosting is another ensemble method that converts multiple poor performing models/trees into a single, stronger prediction model. Cross-validation was used to tune the number of trees included in the final predictive model.

In addition to supporting parameter tuning, cross-validation was used to calculate an average cross-validation (CV) error rate for each of the models. This CV error rate was then used to identify the model that best represented previous voting patterns.

Software and Packages

The analysis conducted within this project was developed using R Statistical Software (R Core Team, 2021) within the RStudio integrated development environment (RStudio Team, 2022).

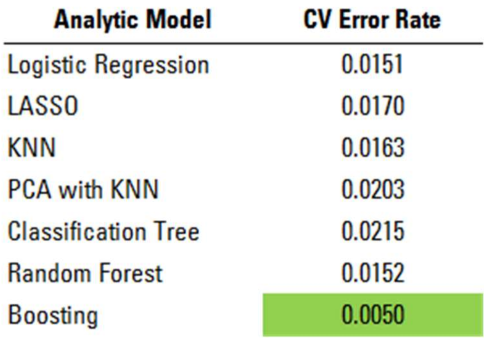
Additionally, code packages were leveraged in creating each of the models as noted below (and cited in the Citations section of the Appendix):



Model	Package	Version
Logistic Regression and PCA	stats	4.1.2
LASSO	glmnet	4.1-3
K Nearest Neighbor	class	7.3-19
ClassificationTree	rpartt	4.1-15
Random Forest	randomForest	4.7-1
Boosting	gbm	2.1.8.1

Analysis and Results

Each of the models were trained and validated, with their respective parameters tuned, using ten-fold cross-validation as described earlier. The following table summarizes the CV error rates for each model. See Figures 1 through 5 in the Appendix for more information on the parameter tuning.



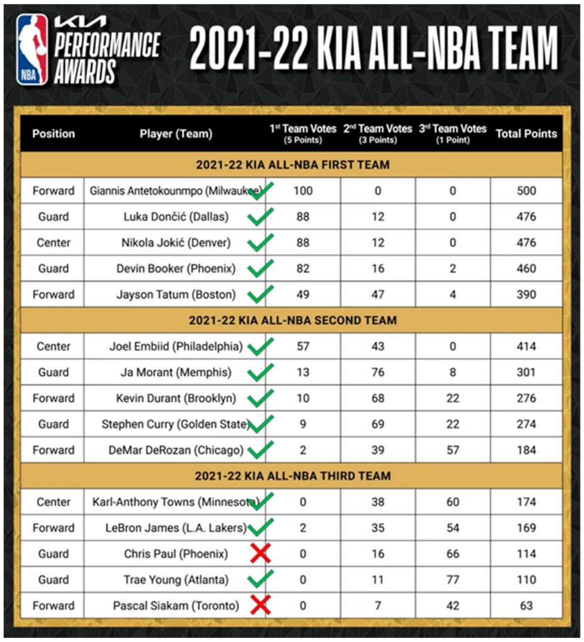
Analytic Model	CV Error Rate
Logistic Regression	0.0151
LASSO	0.0170
KNN	0.0163
PCA with KNN	0.0203
Classification Tree	0.0215
Random Forest	0.0152
Boosting	0.0050

With a CV error rate of 0.50%, **the Boosting model far outperformed the other methods** and was selected as the closest representation of voting in previous years.

This model was then used to assess the actual voting results for the 2021-2022 season, using the prediction data set that had been set aside and not used to train the models.

Notably, the Boosting model agreed with 13 of the 15 All-NBA Team selections. Using the data, however, it is suggested that Chris Paul and Pascal Siakam, who were voted onto the team by the sports reporters, should have been replaced by Rudy Gobert and Jimmy Butler. See Figures 6 and 7 in the appendix for specific probabilities returned by the model.

The following graphic from the NBA has been annotated to illustrate where the selections differ.



KIA PERFORMANCE AWARDS

2021-22 KIA ALL-NBA TEAM

Position	Player (Team)	1st Team Votes (5 Points)	2nd Team Votes (3 Points)	3rd Team Votes (1 Point)	Total Points
2021-22 KIA ALL-NBA FIRST TEAM					
Forward	Giannis Antetokounmpo (Milwaukee) ✓	100	0	0	500
Guard	Luka Dončić (Dallas) ✓	88	12	0	476
Center	Nikola Jokić (Denver) ✓	88	12	0	476
Guard	Devin Booker (Phoenix) ✓	82	16	2	460
Forward	Jayson Tatum (Boston) ✓	49	47	4	390
2021-22 KIA ALL-NBA SECOND TEAM					
Center	Joel Embiid (Philadelphia) ✓	57	43	0	414
Guard	Ja Morant (Memphis) ✓	13	76	8	301
Forward	Kevin Durant (Brooklyn) ✓	10	68	22	276
Guard	Stephen Curry (Golden State) ✓	9	69	22	274
Forward	DeMar DeRozan (Chicago) ✓	2	39	57	184
2021-22 KIA ALL-NBA THIRD TEAM					
Center	Karl-Anthony Towns (Minnesota) ✓	0	38	60	174
Forward	LeBron James (L.A. Lakers) ✓	2	35	54	169
Guard	Chris Paul (Phoenix) ✗	0	16	66	114
Guard	Trae Young (Atlanta) ✓	0	11	77	110
Forward	Pascal Siakam (Toronto) ✗	0	7	42	63

***Source**: NBA.com, 2022 (annotations added)*

Research was conducted to validate the reasonableness of the model's recommendations. While several analysts lamented the omission of one of the two players (Anderson, 2022 and Barnard, 2022 among others), one Hoops Habit article named Gobert and Butler the two top snubs from the 2021-2022 All-NBA Team (Water-Warner, 2022).

Conclusion

While there is solace in having independent industry experts agree with the results of the model created in this analysis, those opinions are also just opinions. For every player that pundits would like to see added to the All-NBA Team, another one would need to be removed. There is certainly no guarantee that those individuals would agree that Paul or Siakam should be the ones replaced on their ideal teams.

Given the wealth of data now available on player performance, it may be time to determine All-NBA Team members based on advanced analytics and not on potentially biased human voting.

The ensemble Boosting method would be a good option because, as a black box approach, it is not directly interpretable; meaning players could not game the system by padding specific statistics. Additionally, as a model trained by previous All-NBA Team voting patterns, it inherently would honor the traditions of the past by incorporating its complex reasoning in a modern way.

Course Lessons Learned

I thought this course delivered a nice balance between theory and the practical applications of several methods and techniques.

I leave the course with a far greater appreciation for cross-validation than I had coming in.

I was VERY appreciative and impressed that Dr. Mei provided interactive office hour sessions throughout the semester. I only had the opportunity to participate "live" once, but the recordings helped clarify some of the course lessons in a way I hadn't initially grasped when watching the formal lessons.

In terms of suggestions or other feedback, I would have preferred a little more clarity on the expectations for the homework assignments. I did well in them but hanging in the air was an uncertainly on whether we should try to create the models as stipulated (which would permit easy review of the results) or if we should take additional action based on things that presented themselves in EDA. TA responses in Piazza to these types of questions appeared to encourage that exploration; but without the clear guidelines on the assignment itself, it was difficult to be confident in how to proceed either way.

Finally, it would be helpful if assignment feedback could be delivered earlier. At times, I received feedback a couple days before the next assignment was due. The homework instructions evolved throughout the semester, implying an expectation of improving work on the student's part. It is critical to have TA feedback early enough to support that improvement.

My most recent example of this is related to feedback from my project presentation. I am adding this and the following paragraph on Saturday April 15 (the day before our written paper is due). I woke up to feedback on my presentation that suggested I shouldn't use the training error to assess my models. I didn't do that; but recognized I needed to do a better job clarifying the use of cross-validation error rates in my deliverables.

I still scored well on the presentation and have incorporated that good feedback into this written report. Thankfully, in my case, that only meant some thoughtful clarifications related to my terminology. For those students or, worse, *teams* that have significant issues, they would now only have a day and a half to determine how to address those items and (re)submit their work. My report had been submitted for over a week at this point; it would have been very frustrating to have to rework a considerable amount of so late in the project timeline.

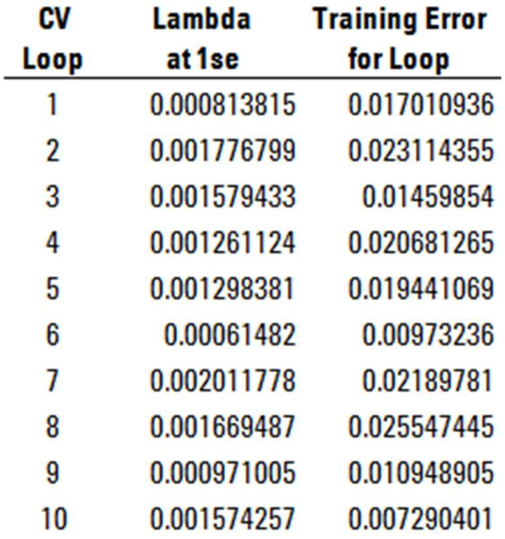
To clarify, I am appreciative of the feedback on my assignments and truly believe those insights are an important element of our growth. I would encourage exploration as to how to get that feedback to students sooner **in relation to upcoming assignments** (not necessarily "quicker" per se). With the TA grading not dependent on the peer feedback students receive, perhaps the work to grade those assignments doesn't have to wait until the peer review window is closed.

All things considered, this has been an excellent course and I'm looking forward to the final exam – the format of which is exciting.

Best regards to Dr. Mei and the team of TAs supporting the course!

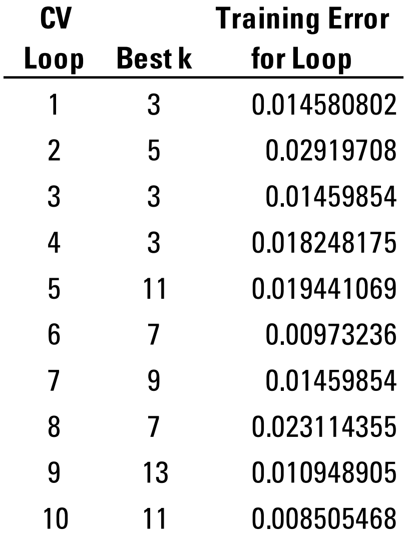
Appendix: Additional Information

Figure 1. Parameter tuning – LASSO. Lambda at 1se for each cross-validation loop.



CV Loop	Lambda at 1se	Training Error for Loop
1	0.000813815	0.017010936
2	0.001776799	0.023114355
3	0.001579433	0.01459854
4	0.001261124	0.020681265
5	0.001298381	0.019441069
6	0.00061482	0.00973236
7	0.002011778	0.02189781
8	0.001669487	0.025547445
9	0.000971005	0.010948905
10	0.001574257	0.007290401

Figure 2. Parameter tuning – KNN. Optimal number of neighbors (k) for each cross-validation loop.



CV Loop	Best k	Training Error for Loop
1	3	0.014580802
2	5	0.02919708
3	3	0.01459854
4	3	0.018248175
5	11	0.019441069
6	7	0.00973236
7	9	0.01459854
8	7	0.023114355
9	13	0.010948905
10	11	0.008505468

Figure 3. Parameter tuning – PCA with KNN. Number of principal components (PCs) cumulatively representing at least 80% of variance in data and optimal number of neighbors (k) for each cross-validation loop.

CV Loop	PCs > 80% Variance	Best k	Training Error for Loop
1	6	3	0.01215067
2	6	3	0.03163017
3	6	3	0.02189781
4	6	3	0.01703163
5	6	3	0.02673147
6	6	3	0.013382
7	6	3	0.02311436
8	6	3	0.03163017
9	6	3	0.01459854
10	6	3	0.0109356

Figure 4. Parameter tuning – Random Forest. Determining the number of variables (mtry) to include and trees (ntree) to create.

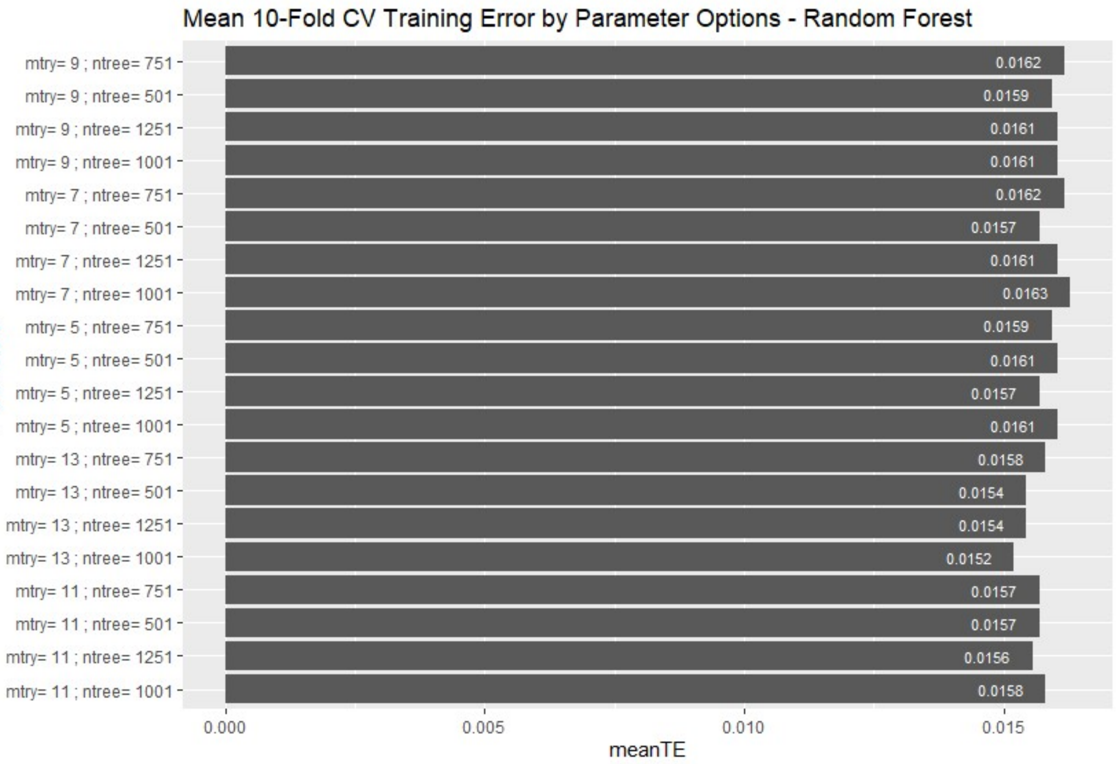


Figure 5. Parameter tuning – Boosting. Determining the optimal number of trees to include in Boosting model.

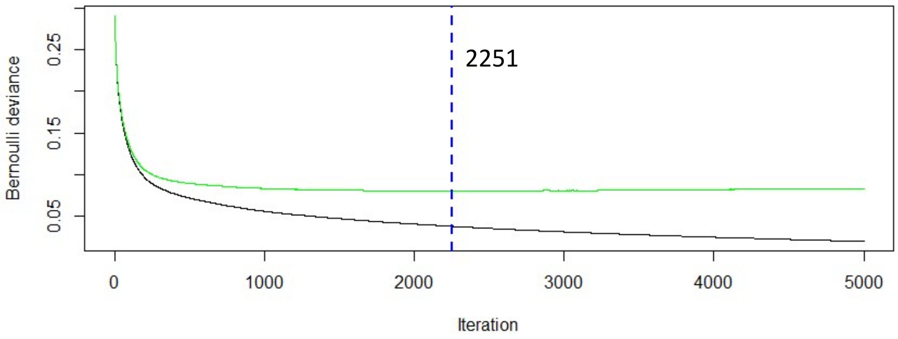
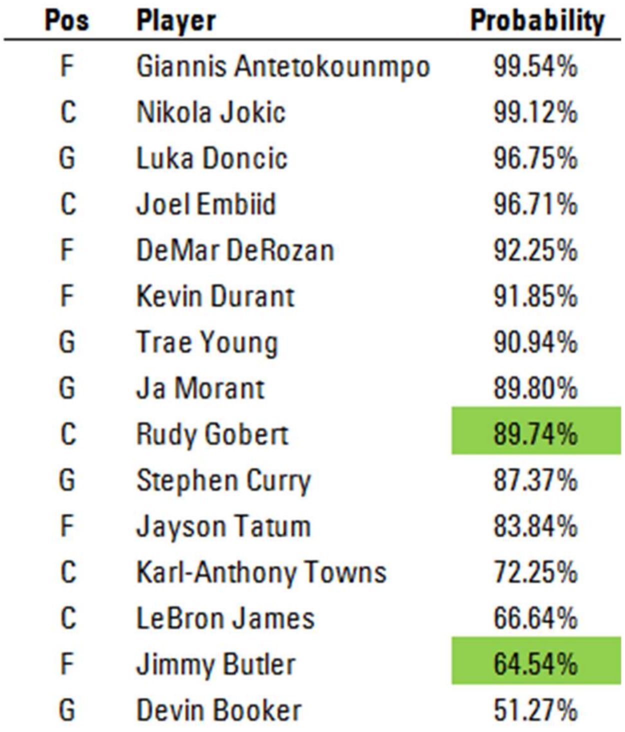
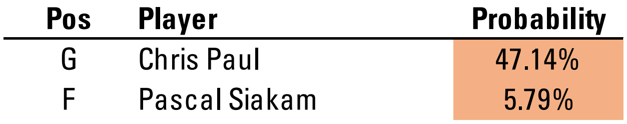


Figure 6. Probability list for players selected to the 2021-2022 All-NBA Team by the Boosting model. Gobert and Butler were not voted onto the actual team by the panel of sports reporters.



Pos	Player	Probability
F	Giannis Antetokounmpo	99.54%
C	Nikola Jokic	99.12%
G	Luka Doncic	96.75%
C	Joel Embiid	96.71%
F	DeMar DeRozan	92.25%
F	Kevin Durant	91.85%
G	Trae Young	90.94%
G	Ja Morant	89.80%
C	Rudy Gobert	89.74%
G	Stephen Curry	87.37%
F	Jayson Tatum	83.84%
C	Karl-Anthony Towns	72.25%
C	LeBron James	66.64%
F	Jimmy Butler	64.54%
G	Devin Booker	51.27%

Figure 7. Probability list for All-NBA Team members excluded from the Boosting model.



Pos	Player	Probability
G	Chris Paul	47.14%
F	Pascal Siakam	5.79%

Appendix: Citations

Anderson, B. (2022, May 25). Gobert, Mitchell left off All-NBA Teams. KSL Sports. Retrieved from https://kslsports.com/486921/gobert-mitchell-left-off-all-nba-teams

Barnard, S. (2022, May 25). 3 Worst Snubs after Release of 2021-22 All-NBA Teams. ClutchPoints. Retrieved from https://clutchpoints.com/3-worst-snubs-after-release-of-2021-22-all-nba-teams

Doll, K. (2022, October 21). Moneyball: The Differences Between the Book & Movie. Shortform Books. Retrieved from https://www.shortform.com/blog/moneyball-book-vs-movie/

FOX Sports. (2016, April 12). The 17 Most Important Records in Sports, from DiMaggio to MJ. FOX Sports. Retrieved from https://www.foxsports.com/stories/other/the-17-most-important-records-in-sports-from-dimaggio-to-mj

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, Vol. 33, pp. 1–22. Retrieved from https://www.jstatsoft.org/v33/i01/

Greenwell, B., Boehmke, B., Cunningham, J., & Developers, G. B. M. (2022). gbm: Generalized Boosted Regression Models. Retrieved from https://CRAN.R-project.org/package=gbm

Helin, K. (2023, March 13). New CBA Reportedly May Feature Games Played Minimum to Qualify for MVP, Other Awards. ProBasketballTalk | NBC Sports. Retrieved from https://nba.nbcsports.com/2023/03/13/new-cba-reportedly-may-feature-games-played-minimum-to-qualify-for-mvp-other-awards/

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, Vol. 2, pp. 18–22. Retrieved from https://CRAN.R-project.org/doc/Rnews/

Merrimack College. (2021, August 24). How NBA Analytics is Changing Basketball. Merrimack College Online. Retrieved from https://online.merrimack.edu/nba-analytics-changing-basketball/

NBA.com. (2022, May 25). Giannis Antetokounmpo, Nikola Jokic, Luka Doncic lead 2021-22 Kia All-NBA 1st Team. NBA.com. Retrieved from https://www.nba.com/news/antetokounmpo-jokic-doncic-2021-22-kia-all-nba-first-team

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

RStudio Team (2022). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

Sports Reference. (n.d.). NBA & ABA All-League Teams. Basketball. Retrieved from https://www.basketball-reference.com/awards/all_league.html

Therneau, T., & Atkinson, B. (2019). rpart: Recursive Partitioning and Regression Trees. Retrieved from https://CRAN.R-project.org/package=rpart

Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with S (Fourth). Fourth. Retrieved from https://www.stats.ox.ac.uk/pub/MASS4/

Vorkunov, M. (2022, March 30). All-NBA Teams are a Great Honor and Also Present Issues for Players and Writers. The Athletic. Retrieved from https://theathletic.com/3214430/2022/03/30/all-nba-teams-are-a-great-honor-they-also-present-issues-for-players-and-writers-basketball-business-digest/

Walter-Warner, H. (2022, May 28). Top 5 Biggest Snubs from the 2021-22 All-NBA Teams. Hoops Habit. Retrieved from https://hoopshabit.com/2022/05/28/snubs-all-nba-teams

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Retrieved from https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2022). dplyr: A Grammar of Data Manipulation. Retrieved from https://CRAN.R-project.org/package=dplyr