Data programming in r final project report

Abhinav Ragupathy, Amar Deep Gautum, Khang Le Tan, Jason Woodson

1. **Background**

Each year a select number of credentialed media members come together to decide who they think were the best players through the entire 82-game regular season: this is known as the All-NBA team. The All-NBA Team is awarded annually to the best players in the league. Originally, All-NBA consisted of two teams, but since 1998 it has been composed of three five-man lineups (1 center, 2 forwards, and 2 guards) that make up the 1st, 2nd, and 3rd team.

What criteria goes into selecting these 15 players? A basketball player’s job on the court is all-encompassing: scoring, rebound, assisting, defense, leadership, etc. Some would argue that media members who have a vote are often influenced by a player’s decision off the court or their sentiment in regard to what their peers think about them. Nonetheless, it goes without saying that several factors go into determining who truly are the best 15 players in each season.

Our group decided to create a model by using players’ regular season statistics to predict which players will be selected to the All-NBA team. Our data was pulled directly from basketball-reference using the nbastatR package which utilizes reputable websites such NBA.com, Basketball Insiders, HoopsHype, and RealGM. The project was completed by using the following additional libraries:

* Tidyverse (opinionated collection of R packages designed for data science)
* Pscl (political science computation laboratory)
* Car (R companion to applied regression)
* Rvest (web page data scraping tool)
* Gbm (implementation of AdaBoost algorithm and gradient boosted machine)
* Caret (set of functions to streamline the process of creating predictive models)
* Rsample (functions to create and summarize different resampling objects)
* data.table (fast aggregation of large data)
* fastDummies (create dummy columns that have categorial variables)
* gt (create presentation-ready display tables)

1. **Data Preprocessing**

We began our preprocessing phase by scraping all traditional and advance statistics from basketball reference for every player in each season from 1980-2021[[1]](#footnote-1). Although we had data available as far back as 1947, we chose to use 1980 as a cutoff since that was the first year that the three-point line was instituted into the NBA and any statistics before that might prove to be inconsistent in any future analysis. The result of the web scraping was a data frame with 17,687 rows and 67 columns to be used as potential feature variables. This data frame will be referred to as df\_all\_seasons.

Next, we again utilized basketball reference and scraped the webpage to get a historic record of the All-NBA teams dating as far back as 1947[[2]](#footnote-2). The resulting data frame required some manipulation (removing empty rows, dropping unwanted columns, renaming columns, etc.), but the result was a comprehensive data frame with every player that had made an All-NBA team as well as the season for which they made the team. This data frame will be referred to as all\_nba.

Our next preprocessing step was to find a way to add an identifier to df\_all\_seasons that indicated whether the player in that season made an All-NBA team or didn’t. To accomplish we created a unique identifier in df\_all\_seasons and all\_nba that was a simple combination of the season and the player’s name[[3]](#footnote-3). We then used a for loop in conjunction with the grepl function to iterate through df\_all\_seasons and check if the unique id in each row was present in the vector of unique ids from all\_nba. The result was either a 1 (yes All-NBA team) or 0 (no All-NBA team) for all players in df\_all\_seasons.

Lastly, we prepped df\_all\_seasons for modeling by removing any columns that would not be used a predictive feature in our modeling process and turned the position column into a dummy variable (center, forward, guard)[[4]](#footnote-4).

1. **Analysis**

For our analysis we chose to run a gradient boosted machine, mainly due to its ability to identify “weak learners” and subsequently attempt to turn them into accurate predictors. Given our dataset had over 50 predictive features, a gradient boosted machine would properly weight the strongest factors while also finding a way to best incorporate features that have less influence on whether a players make an All-NBA team. We first, identified features that our default model deemed to be important by looking at the relative variable influence. The results can be seen below:

Chart

Description automatically generated

It wasn’t much of a surprise that VORP was far and away the most significant variable that the model identified. VORP is a box-score estimate of points per Chart, box and whisker chart

Description automatically generated100 possessions that a player contributes over a replacement-level player (-2) translated to an average team and prorated over an 82-game season. The distribution of VORP for players who made an All-NBA team vs players who didn’t can be seen here. It is expected that the 15 best players in the NBA are well above the contributions of a replacement level player.

Furthermore, we examined the top five most significant variable across five-year increments to validate the trends observed from data based on established observations from NBA.

Table

Description automatically generated

While we did notice some fluctuations across time periods, two things stood out to us: the steady increase in player efficiency across time and the steep drop off in win shares and VORP for All-NBA team members. The decline in win shares and VORP support the notion that the league has gotten deeper over the year and that perhaps winning isn’t as valued in selecting All-NBA players as it once was.

After getting a better understanding of the significant variables that were driving the outcome of our target variable, we went about training our model and tuning the parameters to find our optimal tree size in our gradient boosted machine[[5]](#footnote-5). In the end, we settled on a model with an accuracy, precision, and sensitivity of 0.984,0.803, and .693, respectively, and the following parameters:

* shrinkage = .05 (learning rate)
* interaction.depth = 3 (number of splits the model performs on a tree)
* n.minobsinnode = 5 (minimum number of observations in a node)
* n.trees = 120 (number of trees)
* cv.folds = 10 (number of cross-validation folds to perform)

1. **Predictions**

To test the results of our gradient boosted machine we deployed the model on data from the most recent season that just concluded a few weeks ago. The model did a good job at predicting players at the top end. It is important to note that the top three players the model predicted to make the All-NBA team are the consensus front runners for the Most Valuable Player award this year, further validating our results. However, we did notice a sharp drop off around 77%. While we still expect the model to predict all player’s with at least a probability of .50 to make the All-NBA team, one area of improvement for our model would be to possibly look at lowering that threshold for an even better accuracy rate.

All in all, we are confident in our model’s ability to predict any season’s All-NBA players as seen in the graphic below:

Graphical user interface, table

Description automatically generated with medium confidence

1. **Appendix**

Text

Description automatically generated

**Text

Description automatically generated**

**Text

Description automatically generated**

**4.**

**Text

Description automatically generated**

**5.**

Table

Description automatically generated

1. For loop to aggregate all advance statistics for seasons 1980-2021. [↑](#footnote-ref-1)
2. Scraping basketball reference for All-NBA teams for every season [↑](#footnote-ref-2)
3. Unique identifier [↑](#footnote-ref-3)
4. Creating dummy variables for player positions [↑](#footnote-ref-4)
5. Gradient Boosted Machine training results [↑](#footnote-ref-5)