Methods

Jake Caldwell

10/9/2022

Introduction

The aim of my DA 401 capstone project is to investigate a possible connection between statistical production of NBA players and our familiarity with players. More specifically, if we can recognize an NBA player visually, does that speak to the level of their statistical output. I will be using a mix of observational data (NBA player stats) and experimental data (Survey responses). Because I am testing for correlation between two different types of data, my findings cannot be casual. Instead, only correlation can be discussed. This means that I do not need a baseline measure. My research is unique within basketball statistics, so my findings can launch further projects that scrutinize my post project discussions.

The main structure of my research will involve using machine learning to select variables, surveying subjects to gauge player familiarity, using Ranked Based Overlap to compare lists of players, and creating the statistic to measure player performance. If the feature selection using machine learning cannot be validated reasonably, then an alternative method to doing this project is to create multiple ranked lists using the stats I will explain in this paper as well as creating my own statistic based on research from D. J. Berri (Berri, 1999). Although this might not have the same rigor as using machine learning, it is a plausible backup plan that is easier to validate based on prior statistical research.

Data:

I am using two data sets. One from Basketball-Reference.com and one that will be created from my participant based questionnaire responses.

For the questionnaire responses (Questionnaire will be in the appendix), they will be generated by reaching out to students of Denison University. If these students choose to respond, they will be asked 5 questions that determine each participant's level of expertise and then shown 10 photos of NBA players to see if they recognize them. If they recognize the player, a participant will then be asked a few questions about said player's career. These responses will be recorded as de-identified observations. Once all responses are recorded, the data will be aggregated into counts so that summary statistics and visualizations can be created. This will illustrate the variety and breadth of the sample. Since this is my own data, it will now be cited as (Caldwell, 2022).

The survey data is human data. For this, I have applied for IRB approval and I already have my oral Informed Consent written out as well as assurance of anonymity of their data. Indirect identification is impossible for this data because each survey is de-identified and the data will be aggregated into counts. This means no single observation can be pulled out from the summary statistics. I also accounted for potential bias in this data by using 5 questions at the beginning to determine the level of expertise each participant has regarding the NBA. That way, I can filter results based on expertise level and see how much effect that has on the ability to recognize players. Once I get all survey responses and have aggregated them into counts, I will dispose of the response data. Before that happens, the response data will be stored in a protected folder.

The other data I will be using is per game statistics for each NBA player from 1980-2022 from Basketball-Reference.com (Basketball-Reference.com, 09/21/2022). With about 450 players in the league at a given time, that makes for 18,900 observations. The data includes basic per game statistics such as points per game, assists per game and a few advanced stats like

efficient field goal percentage. This dataset also has a time component: the season, which is a year. Depending on the player, they may have 15 or more seasons with varying stats. This data is open source provided that I cite the day I accessed it and do not sell my work. I can publish this project with this data if I properly give credit to Basketball-Reference.com.

There are many players in this dataset that have only played one season or only play a few games each season. Because of this, I will only be including players who have either played more than one season or who play more than 20 games a season. To clarify, a player may only play 15 games in a season, but this could be because of injury. This means I'll have to check over a player's history of games played. This gets rid of players that might skew the data as well as help the run time of my statistical analysis.

I will calculate two new columns: Player Efficiency Rating and GameScore. These columns are linear combinations of variables already in the data. I will also get rid of rows that do not meet the aforementioned inclusion criteria (see above paragraph). Most of the variables in the dataset will be used in calculating PER and GameScore, making all of the variables integral to my research. The biggest wrangling step for this data is joining all of the seasons into one comprehensive data set. The exported csvs that come from Basketball-Reference.com are tidy to begin with.

Statistical Review

This research involves the use of NBA statistics that carry acronyms and definitions that someone who is not invested in the sport may be unaware of. This section of the methods will explain the general statistic and the reason for its inclusion.

The two statistics that I am going to use for the basis of my feature selection are PER and GameScore. PER stands for Player Efficiency Rating. The equation for PER is:

(1 / MP) * [3P + (2/3) * AST + (2 - factor * (team_AST / team_FG)) * FG + (FT *0.5 * (1 + (1 - (team_AST / team_FG)) + (2/3) * (team_AST / team_FG))) - VOP * TOV - VOP * DRB% * (FGA - FG) - VOP * 0.44 * (0.44 + (0.56 * DRB%)) * (FTA - FT) + VOP * (1 - DRB%) * (TRB - ORB) + VOP * DRB% * ORB + VOP * STL + VOP * DRB% * BLK - PF * ((lg_FT / lg_PF) - 0.44 * (lg_FTA / lg_PF) * VOP) (Page et al., 2013)

The equation for GameScore is:

GS = PTS+.4FG−.7FGA−.4FT +.7ORB+.3DRB+STL+.7AST+ .7BLK −.4PF −TOV (Page et al. 2013)

Below is a table to understand what the statistics make up these two measurements:

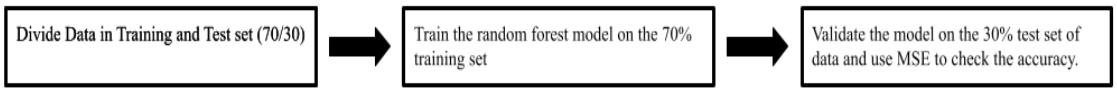
Abbreviation	**Name and Meaning**
MP	Minutes Played - total minutes a player was on the floor
AST	Assists - Number of passes a player made that immediately resulted in a made shot
FG	Field Goals - Number of made shots in a game
FT	Free Throws- Number of made free throws in a game
VOP	Value of possession - average points per possession
TOV	Turnovers - number of times the player lost the ball to the other team
FGA	Field Goals Attempted - total number of shots taken by a player (make or miss)
DRB	Defensive Rebounds- number of rebounds grabbed by a player while on defense
ORB	Offensive Rebounds - number of rebounds grabbed by a player while on offense

FTA	Free Throws Attempted - Number of free throws taken (made or missed)
STL	Steals - number of times the player stole the ball from the other team
BLK	Blocks - number of times the player blocked the shot of an opposing player
PF	Personal Fouls - Number of fouls the player committed
TRB	Total Rebounds - DRB + ORB
PTS	Points - Total points scored by a player during a game

PER is a measure of player contribution to their team on a per-minute basis. GameScore is a measure of how well the player performed based on a linear combination of their basic in game stats. There are a lot of variables to understand, but these components are all basic in game stats that one can see happen with their own eyes. When I run my machine learning model to perform feature selection, I will use PER and GameScore as my dependent variables so that I can see which of the independent variables from the above table matters the most.

Feature Selection

I am performing feature selection using PER and GameScore so I can create a hybrid statistical model that contains the most influential features of each statistic. I plan to write a random forest algorithm from the randomforest package version 4.7-1.1 (Liaw et al., 2002) using R version 4.1.2 (2021-11-01). Below is a diagram that represents how the feature selection process would work:



This will repeat for both PER and GameScore. However, once both models have been run and their tree diagrams have been created, there is a second validation method. The best PERs are around low to mid 30s and the best GameScores are around mid to upper 50s. If the stat that I produce doesn't fall within this range for the best seasonal performance, then I built the statistic incorrectly.

Human Subject Study Design

In order to generate the second ranked list for my final comparison, I am using human subjects research to determine the recognizability of NBA players who were selected to the NBA 75th anniversary team (Only players who played after 1980 are included). The basic outline of this study is as follows:

1. Ask each participant 5 questions to determine their level of NBA knowledge. (Desmarais et al., 1995).
2. Show them 10 faces randomly selected from the 47 players in the NBA 75 team (47 players played in 1980 or later).
    a. If they recognize a player, ask them a few questions about the player's career
3. Debriefing questions about how they recognized those players.

The full questionnaire will be in the appendix of this paper.

This study design will allow me to test for familiarity of NBA players because I can generate a count of the number of times a player was recognized. The list will be ordered greatest to least to show the most recognizable players. The extra questions asked will help facilitate the discussion section of my research. I can use the answers to the subsequent questions to explore possible connections between statistical production and recognizability of NBA players.

Ranked Based Comparison

After both ranked lists have been generated, the last step in this project is to compare the lists to see if they are similar enough to say that familiarity is correlated with statistical production. The statistical test required for this is Ranked Based Overlap (RBO) (Joshi, 2021). RBO allows me to compare the similarity of two ranked lists of the same size. The output is a similarity measure (Bounded between 0,1). A score of 0 means disjoint lists and a score of 1 means identical lists. RBO also allows for a weighted valuation of the top n ranks in a list. For example, if I have a list of 10 items, I can choose p=0.6 to weigh the top 3 items more heavily. The advantage that this has for my project is that if the lists I output are dis-similar after the top 20, I can weigh the top 20 more heavily to see if they are more than 50% similar. With that being said, my significance level will be set at .5. That means that if the RBO score is greater than .5, then there might be some correlation between familiarity and statistical production. This will be done using the TopKLists package (Version 1.0.8) in R (Schimek et al., 2022) in R version 4.7-1.1.

Appendix:

Questionnaire:

1. **Level of Interest:**
    a. What does the NBA stand for?
    b. What does PPG stand for?
    c. Who is the all time leading scorer in the NBA (Regular Season)?
    d. What does PER stand for?
    e. Who is the last player to win 6th man of the year?
2. **Player photos**
    a. If they recognize the player shown, participant will be asked the following:
        i. What is the player's career statline (Points per game, Assists per game, Rebounds per game)?
        ii. What team has the player most recently played for?
        iii. Are you a fan of this player?
3. **Debriefing Question**
    a. Do you think that you recognized the players that you did because of media coverage?

Bibliography

Basketball-Reference.com - Basketball Statistics and History. https://www.basketball-reference.com/. (09/21/2022)

Berri, D. J. (1999). Who Is "Most Valuable"? Measuring the Player's Production of Wins in the National Basketball Association. *Managerial and Decision Economics*, *20*(8), 411–427. http://www.jstor.org/stable/3108257

Desmarais, M. C., Maluf, A., & Liu, J. (1995). User-expertise modeling with empirically derived probabilistic implication networks. *User modeling and user-adapted interaction*, *5*(3), 283-315.

Joshi, P. (2021, January 12). *RBO V/s Kendall Tau to compare ranked lists of items*. Medium. Retrieved September 29, 2022, from https://towardsdatascience.com/rbo-v-s-kendall-tau-to-compare-ranked-lists-of-items-8776c5182899

A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Page, G. L., Barney, B. J., & McGuire, A. T. (2013). Effect of position, usage rate, and per game minutes played on nba player production curves. *Journal of Quantitative Analysis in Sports*, *9*(4), 337-345.

Schimek, Budinska, Jie Ding, Kugler, Svendova, Lin and Pfeifer (2022). TopKLists: Inference, Aggregation and Visualization for Top-K Ranked Lists. R package version 1.0.8. https://CRAN.R-project.org/package=TopKLists