Determining Archetypes of NBA Players

Spencer Dooley
STAT 287

spencer.dooley@uvm.edu

ABSTRACT

The data used in my project was aimed to represent how a player plays the game (attempting to capture the players playstyle). Four data sets were used, one consisting of traditional stats, one of advanced metrics, one with shot location data, and one with touch data. Overall, I ended up with seventy-four features, so clearly I needed to reduce dimensions. Using Principal Component Analysis (PCA), I reduced my data set to twelve features (components) while still retaining 86% of the original information. After reducing dimensionality, K-Means clustering was performed to cluster the players into ten clusters. Once these clusters were determined, all that was left to analyze these clusters in order to determine what subgroup of players were being represented. Using a "One vs. Rest" approach, I broke down a multi-class classification problem into ten binary classification problems. These ten classification problems were easily modeled with Logistic Regression. After this multi-class-classification, the feature importance in determining each cluster was easily available which made the actual cluster analysis much more straightforward, and thus the subgroups were identified.

Keywords

Dimensionality reduction, Clustering, NBA Prediction, K-Means, Principal Component Analysis, NBA Archetypes

1. INTRODUCTION

Attempting to predict the future is something involved in almost any decision people make. In the world of sports, people have been trying to make accurate predictions since the first ball was thrown. The techniques and models used and analyzed in this project has the potential to take us one step closer to more accurately predicting NBA games. NBA prediction models have become increasingly complex as there are more and more advanced metrics. However, one aspect of the game lots of models fail to include, is that of how players interact within a game, and how their respective playstyle is modeled. In this project, I clustered players using statistics that I thought would represent how a player played and interacted within a game. I hoped to then be able to cluster players by similar playstyles and learn more about the NBA overall in the process. From this information, we can learn more about the makeup of NBA teams and, if more historical data is incorporated, could learn more about the shift in playstyles over time. The more we can learn about how the players in the NBA interact within a game, the more NBA front offices will be able to improve their teams.

2. RELATED WORK

In a project done by Stanford student Muthu Alagappan, a similar concept was explored. Alagappan used topological data analysis to form groups of players (represented by color-coded nodes) which were connected by edges representing statistical similarities between players. In this analysis, 13 different groups were determined. However, while some of these groups began to explore the pure player playstyle, others seemed more like the role (playing time being a main factor) the player had on a team. While the analysis is similar to what I am aiming to do, the difference will be found in the data used. Alagappan followed the form of other similar works and used mainly just some traditional statistics (points, rebounds, assists, etc.) and some slightly more advanced metrics (usage, PER, etc.). I do not believe these statistics alone necessarily determine a player's playstyle. What I mean by this can be conveyed easier with a quick example. Let's say there are two players, one who scores the majority of their points close to the basket, and one who scores most of their points from behind the three-point line. Now, each of these players may average twenty-three points per game, yet they score those points in opposite ways. What I want to incorporate into my model, which related works always seem to leave out, is statistics that show *how* the player scores, or *how* the player gets their assists.

In another related work, done on *Towards Data Science*, more frequency-based stats were used in determining player clusters. This project is on the opposite end of the spectrum as Muthu's, as this one did not really take role into account at all. For example, one player could be a star on a team, and another seen as a bench player, however, if they have similar frequencies of types of plays, they would be grouped together.

I am looking to be somewhere in-between these projects. I really do want to encapsulate the players true playstyle, but also consider the influence (comes more so from a player's role) a player has on a game.

3. METHODS

3.1 Data Collection, Cleaning, EDA

I viewed the data collection aspect of this project as one of the most important parts. Deciding what data to use was certainly a balancing act of getting data that represented a player's playstyle, but also maintained the importance of player role. With this in mind, I decided to web scrape four data sets from NBA.com. Two of the data sets were traditional and advanced statistics – I hoped these statistics would represent the role a player had on the team. The other two data sets were *shooting location data* and *touch data*. The shooting location data gave me the information of what distance from the hoop certain players score from. The touch data provided similar information as the frequency data discussed in [2] – the amount of touches a player gets at certain spots on the court, number of dribbles taken, time of possession, etc. The latter two data sets was included to give me insight on the player's true playstyle.

As far as data cleaning and EDA, I did a few things to remove unnecessary data points and features as well as fill in missing values. I began by removing players who did not play in at least fifty percent of their team's games or didn't play at least fifteen minutes per game. I used a heat map to determine if any features were especially redundant, which many of them were. I was not surprised by thins and removed many of the very redundant ones. These features were typically features that were calculations done (like field goal percentage). With over 80 features, I decided to drop many of these percentage type features as they did not provide me with much extra information. As for missing values, there were not many, and the values that were missing were filled with the value of zero. I decided to fill them with zero simply because if the value was missing in the first place, it was implied the player did not perform a certain action enough to have a large enough sample size for an accurate statistic. Thus, zero was a reasonable value.

3.2 Dimensionality Reduction

With 66 features (excluding those such as age, wins, losses, etc.), I used Principal Component Analysis to reduce dimensions. I

wanted to retain at least 85 percent of the variance from the original data. I was able to do this by including the top twelve principal components as the features in determining clusters. I chose to use PCA as interpretability was not important to me, as I was simple using the components to cluster players, then by merging the cluster information back with the original data, I'd have all the information I needed.

3.3 Clustering

After reducing dimensions, I then had to cluster the players based on the similarity on the twelve principal components. I decided on using K-Means clustering for this. Although, in the future I'd like to use a more complex form of clustering or other method to group similar players more accurately (mainly because K-Means doesn't cluster outliers well – outliers are necessary in this kind of data), K-Means was easily applicable here and easily understandable. I used silhouette scores for clusters ten to fifteen to determine the ideal number of clusters, k, for k = 10 to k = 15. The max silhouette score came at k = 3, however I wanted to have at least ten different clusters.

3.4 Determining Feature Importance

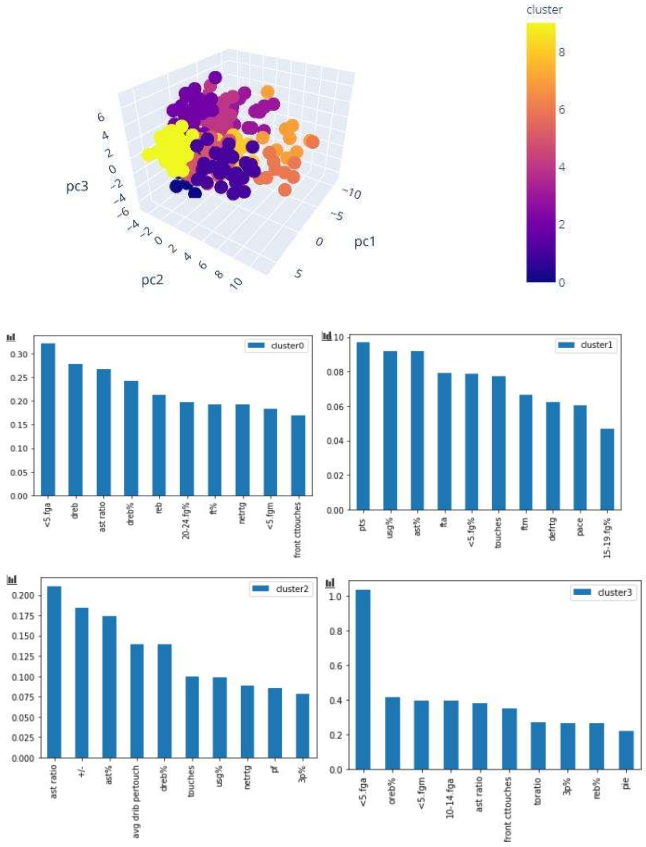
The biggest roadblock in this project was determining how to efficiently and, somewhat scientifically, determine accurate names for the clusters. After some careful consideration and research, I decided to use a form of Multi-Class classification. I used a "One vs. Rest" approach which allowed me to use Logistic Regression classification. I essentially made ten different Logistic Regression binary classification models, one for each cluster, which classified players as either *in cluster x* or *not in cluster x*. By doing this, I could then easily extract feature importance for determining each cluster.

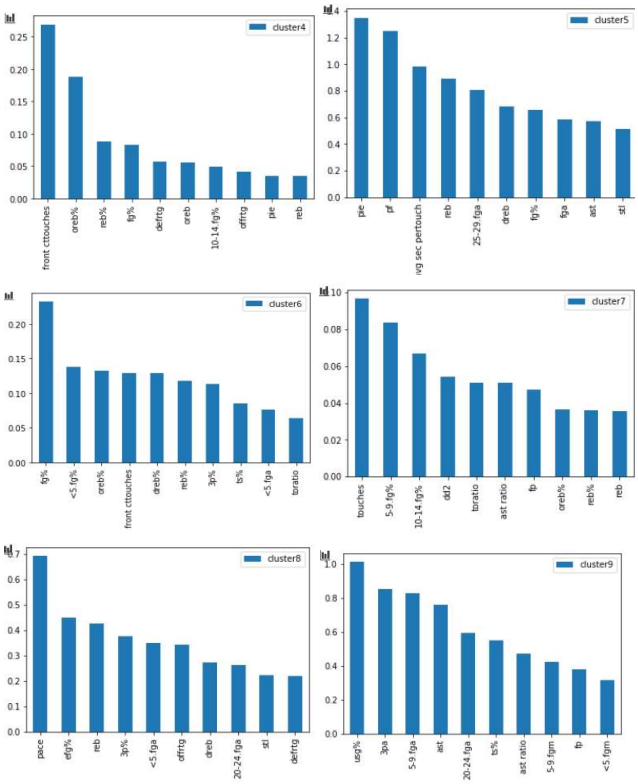
4. RESULTS

Using the top features in determining each cluster and looking at the players closest in Euclidean distance from their respective cluster's center point, I was able to come up with reasonably accurate titles for ten different player archetypes. Below are the clusters along with some defining characteristics:

1. Cluster 0: High Level 3 and D
- Key characteristics: Great outside scoring, good defensive ability, ability to score on drives
2. Cluster 1: Scorers/Multi-Dimensional Players
- Key characteristics: High scoring output, high usage, lot of touches
3. Cluster 2: Backup Point Guards
- Key characteristics: High assist ratio, above average number of dribbles per touch, solid three-point shooting
4. Cluster 3: Athletic Wings
- Key characteristics: Inefficient overall, inefficient shooting, high pace of play
5. Cluster 4: Dunk and Defense Post Players
- Key characteristics: Majority front court touches, high rebounding, good defense, scoring mainly within 14 feet of basket
6. Cluster 5: Multi-Dimensional Role Players
- Key characteristics: Good scoring, Average number of rebounds/assists, multi-faceted scoring, can tend to be inefficient
7. Cluster 6: Backup Bigs
- Key characteristics: Majority front court touches, interior scoring (dunks), high number of rebounds
8. Cluster 7: Skilled Bigs
- Key characteristics: Lots of touches, high scoring, average to below average defense, high number of assists relative to post players
9. Cluster 8: 3 and D (mostly 3)
- Key characteristics: Play at a fast pace, mainly scoring from behind three-point line, decent defense
10. Cluster 9: Stretch Forwards
- Key characteristics: Outside and inside scoring ability, high rebounding, good shooting stats relative to position

4.1 Models





4.1.1 Model Details

The first, three-dimensional, graph shows the clusters of players when using only 3 principal components. Using only 3 components only retained about 56 percent of the variance from the original data, however, I still thought the graph did a good job of showing what was happening in a higher dimension. Even from the 3-D version, some clear clusters start to become apparent.

The following ten bar charts show the top ten most important features used in determining each cluster (0-9).

5. DISCUSSION

Going forward, I am going to consider not using advanced metrics in determining players playstyle. Advanced metrics do a great job of analyzing how efficiently a player plays the game, but it does not necessarily show *how* a player plays the game. The resulting clusters were on the right track of what I had in mind, and for the most part the cluster make sense. However, I noticed some of the advanced metrics were highly important determining features. I would like the advanced, efficiency-based stats to take a back seat in determining a player's archetype. What is very promising though from these results is by simply using these four data sets, a simple clustering algorithm, and some simple dimensionality reduction, the clustering model was quite accurate. Now, with even more knowledge of what data I should use, it will be easier to continue fine tuning this project until the output is something anyone involved in the NBA could use.

6. FUTURE WORK

Overall, the most exciting things to come out of this project for me were the endless possibilities I began to envision to take this project farther. For starters, once the clustering is fine tuned to a point I am satisfied with, I can have a geographic representation of player playstyles. Do certain basketball hotspots (New York, Los Angeles, Chicago, the Midwest, etc.) have higher proportion of different playstyles? There is so much we could learn from that type of visualization. Another step farther is incorporating a binary "all-star" variable to see which group of players has the highest proportion of all-star. Since all-star status is determined in part by voting, is there a certain playstyle that 'excites' fans more? Perhaps what is most exciting, and what has the potential to turn this into a true 'product', is the potential to use this kind of information to simulate individual games, and thus full NBA seasons more accurately. Understanding the way a player plays a game is the first step to modeling how a player interacts within a single game. Using the frequency-based statistics, the shooting location data, and incorporating a few other data sets representing other aspects of a player's game, probabilities of a player performing any action within a moment during a game (make/miss shot from certain spot, pass for/not for assist from certain spot, turn ball over, etc.) can be calculated. From this information, a very detailed simulation of a single game can be done thousands of times, which can then be done over and over again for each game during the course of a season to accurately predict team's future success. The first step in this direction is modeling a player's interaction within a single game, which I modeled by determining player playstyles.

7. CONCLUSION

To wrap things up, let us remember what we were investigating. I set out with the goal of determining NBA player playstyles present in the current state of the NBA. I aimed to make the clusters as representative as possible of how the player plays and how the player acts within a single game. The current results make huge, confusing data sets of NBA statistics easily interpretable to any fan of the game. Instead of trying to analyze hundreds of features by the eye test, a player's playstyle is simply determined by all these features, then presented to the user in a straightforward way. What is most exciting, as alluded to previously, is the future implications of this type of data analysis. By modeling a player's individual actions within a game, based on their playstyle, both simulating and, in turn, predicting NBA games accurately becomes a reality.

8. REFERENCES

[1] Alagappan, M. (n.d.). *Redefining NBA Positions* [Scholarly project]. In *Muthu Alagappan*. Retrieved December 06, 2020, from https://muthualagappan.org/

[2] James. (2019, October 07). Clustering NBA Playstyles Using Machine Learning. Retrieved December 06, 2020, from https://towardsdatascience.com/clustering-nba-playstyles-using-machine-learning-8c7e8e23c9