NBA "Moneyball": An Analytical Approach to Cost-Effective Team Formation Using Hierarchical Clustering and Regression

Calvin Yu, Daniel Lisko, Toan Vang, Ruchi Bhavsar

Abstract– Advanced statistics in sports drive the selection of the best players, which lead to an increased number of wins and improves fan support in addition to the overall economic growth for their respective cities. Achieving a well-rounded team is often easy for already successful and "rich" teams given their power and resources. However, a majority of teams are less successful in comparison to the top teams. This trait often means limited salary caps and a challenging journey out of the bottom and middle tier. By finding a way to generate these teams on budgets, we hypothesize that there would be increased competition and overall viewership of the sport. We devised a model that is capable of generating teams given a salary cap using linear regression and a method of capturing a set of player's total win potential given their advanced statistics. To preface our initial goal, we also performed hierarchical clustering to determine if a player's success and performance is attributed to his advanced statistics. Our metrics show that the teams formed are well within a budget and are viable teams based on historical performance.

1. Introduction

By taking a look at the case study of the general manager of the Oakland Athletic's, a baseball team in Major League Baseball, we see success in a sports team as a direct result of player analytics despite limited salary caps. Although their team fell short of the ultimate goal of winning the World Series, the period of relative success led to a revival in the team's adoration and allowed for continued growth in a seemingly hopeless team. We intend on solving this problem with data mining and machine learning techniques in the field of professional basketball.

Solving this problem has both economic and social implications. In general, we can infer that having better players leads to more wins and success as a team. As a result, fan culture and viewership increases, allowing for economic growth in cities. By emphasizing the creation of teams with limited salary caps, competition would increase and similarly increase viewership. Data mining and machine learning techniques not only provide a more automated way of finding these players but also reduces the need for domain knowledge when analyzing player statistics. Furthermore, such techniques could be applied to any team sport and allow for similar benefits as previously mentioned.

2. Background

The NBA is going under a massive revolution and only a few teams had an analytics department a decade ago. Today nearly all of the teams have data analysts for either spotting undervalued players, choosing team's players or more. Ami and Mochamad [2] analyzed the best

selling book written by Micheal Lewis - Moneyball [1]. The article talks about how Billy Beane, a baseball team general manager, came up with a radical action; he changed the method of acquiring talents for the team. He, along with his assistant, used statistical data to analyze and choose the team's players which gave excellent results for his team.

D. Thenmozhi *et al.* [3] covers another sport, cricket, that benefits from using data driven solutions to evaluate whether a team can win an ongoing match. They were able to achieve the prediction accuracy for each home team in the IPL(Indian Premier League). To use mixed data in analysis Pierpaolo D'Urso *et al.* [4] proposed a robust fuzzy clustering model. The clustering method combines the dissimilarity matrices with weights during the optimization process for every attribute. This clustering method would help in finding clusters that would have been otherwise hidden unless a multi-attribute approach was used.

Luca Pappalardo *et al.* [5] used a data-driven framework, PlayerRank, that offers a multi-dimensional and role-aware evaluation of the performance of the soccer players. After comparing the framework to other known algorithms for performance evaluation along with the player evaluations by professional scouts, PlayerRank is found to perform better. Nguyen Hoang *et al.* [6] used machine learning and deep learning to predict the NBA players' performance and popularity. It was indicated in the results that the most important and contributing factor for the players in any team is their scores.

3. Methodology

3.1 Overview

To determine if advanced statistics are correlated with player performance, we first perform hierarchical clustering. If certain tiers of players begin to form as result of the clustering, then we can further justify our assumption of using statistics to create viable teams. Next, we attempt to generate viable teams with a limited salary cap using advanced statistics, performance of teams in the 2021-22 NBA season, and a linear regression model. To do so, we predict the number of wins that a set of players could produce, choose the team that outputs the most wins, and compare the predicted player's average advanced statistics against current teams.

3.2 Unsupervised Player Grouping

3.2.1 Data Collection and Aggregation

BeautifulSoup (https://www.crummy.com/software/BeautifulSoup/) was used to collect advanced player statistics from basketball.realgm.com for each player from the 2010-2022 NBA seasons. The median statistics were calculated for all the players who played during that time. We parsed https://en.wikipedia.org/wiki/NBA_All-Star_Game to get the number of All-Star

appearances for all the players. Players' names not found on the Wikipedia page were considered to have not played in any All-Star games.

3.2.2 Hierarchical Clustering

Using hierarchical clustering, we will group statistically similar players in the NBA based on how they are clustered. We will then determine the tier rank for each group by identifying advanced statistical characteristics of each group. For example, group A might have more impact players than group B, which means group A will be tier 1 and group B will be tier 2.

3.3 Team Prediction

3.3.1 Data Collection and Aggregation

Data from www.basketball-reference.com was collected and aggregated to form the necessary dataset for our prediction and clustering tasks. This data included the advanced statistics of players currently active, their respective salary, and the win/loss records of each team all from the 2021-22 NBA season. We then filter the player data based on certain requirements such as playing a minimum number of games and a minimum number of minutes.

3.3.2 Generating Training Data

We first need to generate a train set for our regression model. The training data is based on the 2021-22 season standings and the players of their respective teams. Given a list of player statistics and salaries, we can iterate each player, determine the team they play for and filter the ones that do not fit the minutes and games played requirement. Then we can group players according to team and generate combinations of teams. Lastly, we average each team’s players’ advanced statistics to obtain the features or independent variables and we assign the number of wins from their teams in the 2021-22 season standings

3.3.3 Generating Testing Data

The testing data is not from the same distribution as the training data since we want to generate teams that haven’t been formed before and for which we don’t have historical data regarding wins. To generate sets of n-person teams, we randomly sample unique combinations. However, since the number of combinations sampled out of the relatively large pool of players can be overwhelming, we simply limit the number of combinations. We also ensure that each position of a team is filled (small forward, power forward, center, point guard, and shooting guard) and that the combined salaries are between a range. Even though this filtering decreases the variations in players further, it provides a higher chance of getting viable teams that are not strictly based on statistics. Once we have sets of n-person teams, we can average their advanced statistics and begin training.

3.3.4 Regression Model

A simple linear regression model is fitted on the training data to predict the number of wins produced by each team in the test set. The predictions are sorted in descending order based on the number of wins and we can obtain the top-k teams based on the number of wins.

4. Code Details

4.1 Cluster model

For this analysis, we will collect player stats from the 2010 - 2022 seasons and then average those stats for each player. For the clustering model, we will use advanced players' stats since many of these advanced statistics consider a player's contribution to a team's success. Since there are 20 features to consider, we first perform a principal components analysis (PCA) for dimensionality reduction and then select the components with the highest explained variance. Before performing the PCA, we will also perform standard scaling so that all the features have the same weight when constructing the PCA.

Since we want to group players with similar feature characteristics, we will perform hierarchical clustering to group players into different tier rankings (i.e., tier 1 being the best players and tier 5 bench players). First, selecting the top components, we will evaluate several clustering techniques(single linkage, complete linkage, simple average, and Ward's method) and then select the best method that fits our data - based on the cophenetic correlation coefficient.
To further evaluate the selected method, we will visually inspect the hierarchical tree to evaluate cluster separation. Additionally, we will look to see if each group has differences in advanced stats. For example, we expect a higher-ranking group to produce more all-star players with similar skill sets. Finally, we will look at the top 20 players , based on player efficiency ratings, of each group to see if the grouping of players makes sense. For example, top-tier player
should have the most significant impact on a game

4.2 Regression model

After acquiring the data, we will generate training data. This data will consist of teams of players and their averaged statistics as the independent variables and that team's number of wins in the 2021-22 season. We will reduce the size of teams to say, 5, to allow for different combinations of players within the same team and an increase in our training data size. This can be simply done with the Python combinations library.

We will then generate a test set by picking random players and averaging their statistics. The reason for choosing random distinct players to form a team is that generating all possible 5 person combinations out of all the possible players takes a significant amount of time. Even with a subset of the players chosen, there are still 11 billion possible teams. This set will be different

from a typical test set in a machine learning setting since it is not derived from the train set. We do this to make random teams that we have not seen before and do not have win labels for.

Lastly, we will train and apply a regression model from the Sci-kit Learn library to produce the number of wins each random team has and select the highest output as our "best" team. To evaluate performance, we can compare the averaged statistics of the predicted teams to currently successful teams. Additionally, we can see how many players in that team are from top 10 teams in the 2021-22 season and how many are in the bottom 10. This simply requires some Pandas functions to filter the rows.

5. Results

5.1 Unsupervised Clustering Results

Prior to performing hierarchical clustering, we first compressed our dataset into a lower dimensional space by performing principal component analysis (PCA). We selected the first two principal components, PC1 and PC2, because they contribute to the most explained variance (54.4%) of the data set (Supplemental Figure 1).

Using the first two principal components, we evaluated the clustering performance of several clustering methods (single linkage, complete linkage, simple average, and Ward's method). Complete linkage, with a cophenetic correlation coefficient of 82%, was chosen for clustering NBA players (Supplemental Figure 2).

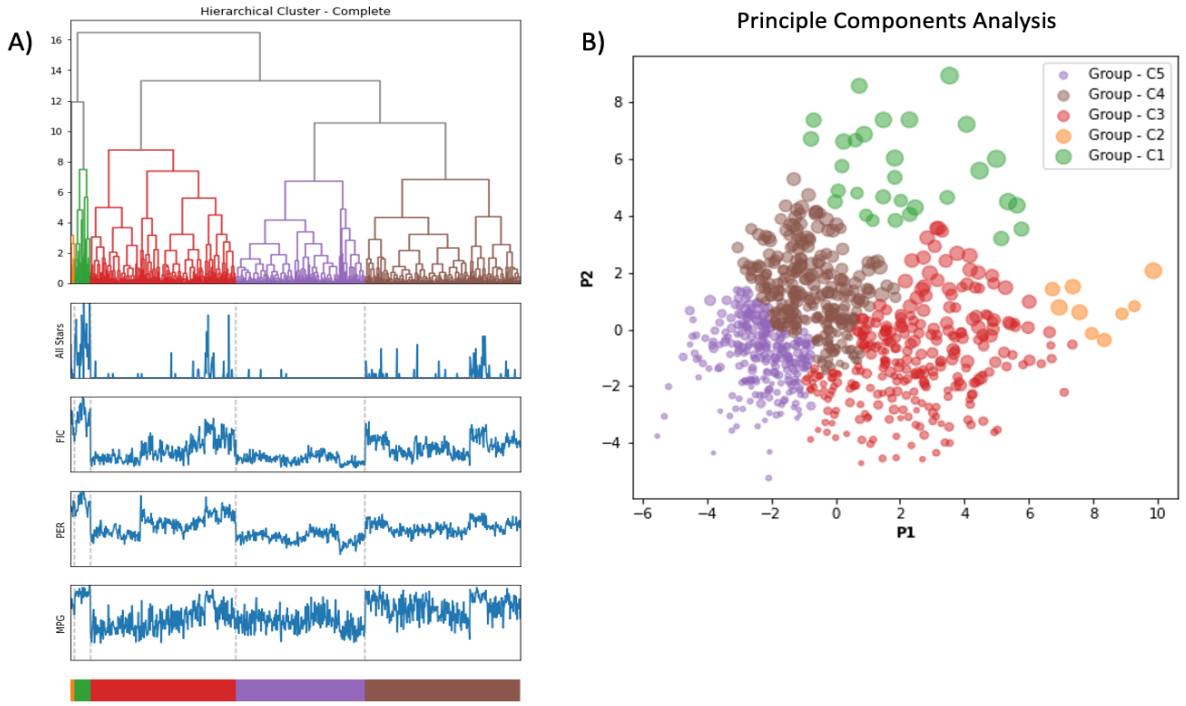

Figure 1. Hierarchical clustering analysis of NBA players by complete linkage using average advanced statistics spanning the 2010 - 2022 seasons. A dendogram (A) Shows the clustering of NBA players and the number of All-Star games, FIC, and MPG for each player. (B) A principle components analysis (PCA) further demonstrates the clustering of players in a 2-D space - the size of dots relates to a players PER.

Using complete linkage (Figure 1), we see that there are five types of players in the NBA and that players within clusters share similar characteristics. For example, players from group C1 and C2 tend to be higher caliber players. Players from this group typically play more minutes per game (MPG), contribute to a team's success (PER and FIC) and have more All-Star players. Differences are observed between each group when comparing all the advanced statistics (Figure 2). Group C1 has a higher AST%, FIC, PER, MPG, USG% and number of All-Star players. Group C2 leads in eFG%, ORB%, DRB%, TRB%, BLK%, PPS, ORtg, and eDiff. Group C5, were the lowest in most categories, especially in PER, MPG, FIC, and number All-Star players. However, they were one of the highest groups for DRtg.

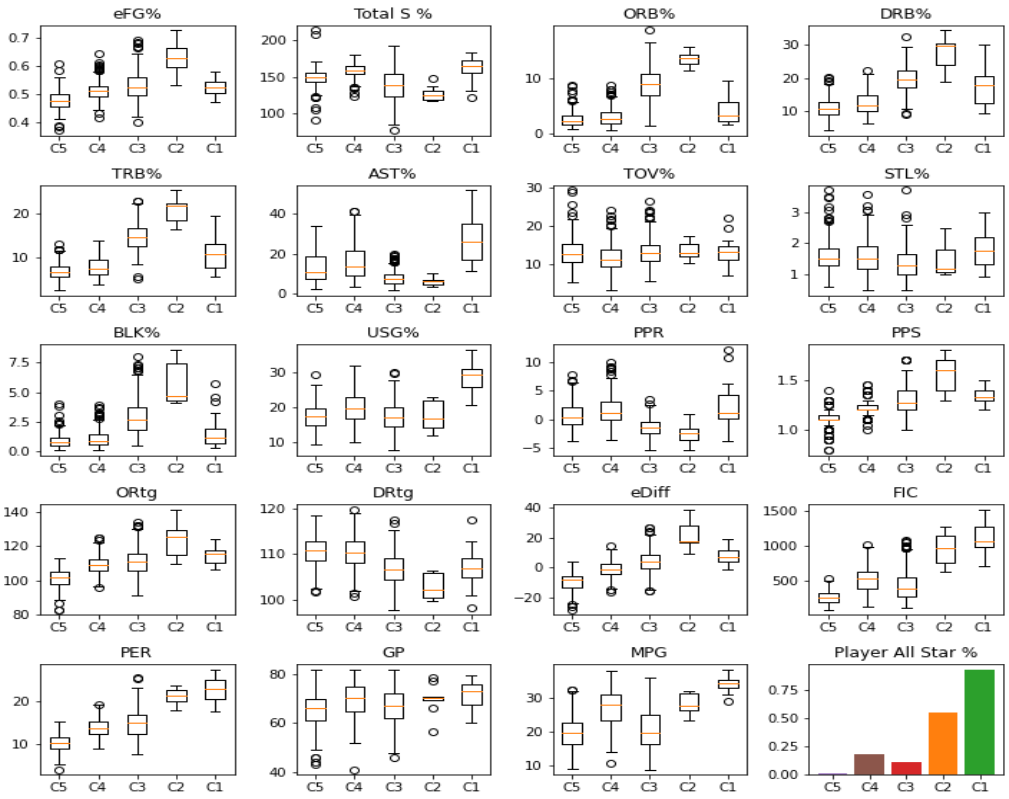

Figure 2. Comparison of advanced statistics between each cluster of NBA players that were generated by hierarchical clustering. Player All-Star % is the percentage based on the number of All-Stars within a group. *See supplemental Table 1 for description of each advanced statistic.*

Figure 3. Top 20 players, based on PER, for each cluster that was generated by hierarchical clustering.

5.2 Team Prediction Results

The following teams were predicted with the individual and averaged advanced statistics along with the salaries. To get a better understanding of the players chosen, we outputted the top 5 teams based on their number of projected wins in order. These teams are ranked from 1-5 in our results section. For comparison purposes, the last team based on the fewest number of projected wins is also determined.

After generating these teams, we can compare the advanced statistics to those of currently successful teams. The teams chosen for comparison were the Phoenix Suns, who had the best win/loss record in the Western Conference, the Miami Heat, who had the best win/loss record in the Eastern Conference and the defending NBA champions, the Golden State Warriors. The advanced statistics were categorized according to their indication of a team's viability. For example, PER is the per minute productivity. This stat, when averaged across a set of players, should indicate a viable team if higher. On the other hand, TOV% is the number of turnovers per 100 plays, which is ideally lower. Lastly, there are stats such as win shares that do not indicate a team's viability when averaged. These stats are used to compare players with one another since they deal with team contributions. The table in (Supplemental Table 1) gives a full description of how we grouped these stats.

Additionally, we can determine what percentage of players in the generated teams were on the top-10 or bottom-10 teams during the 2021-22 season. We would expect a higher percentage of the predicted team to be a part of current top teams with a low number of bottom 10 teams. The reverse is expected for the last ranked predicted team.

Size of Team	Position Matters	Min. Games	Min. Minutes	Salary Range (min)	Salary Range (max)
5	Yes	41	25	$50,000,000	$70,000,000

Rank	Team	Projected Number of Wins	Average 2021-22 Season Salary
1	LaMelo Ball, Kent Bazemore, Devin Booker, Jarred Vanderbilt, Robert Williams	60	$50,164,532
2	Bruce Brown, Damion Lee, Royce O'Neale, Bobby Portis, Pascal Siakam	60	$50,993,133
3	LaMelo Ball, John Konchar, Kawhi Leonard, Kelly Olynk, Dean Wade	60	$63,752,077
4	Jimmy Butler, Robert Covington, T.J. McConnell, Bobby Portis, Anfernee Simons	59	$64,778,089
5	DeAndre Ayton, LaMelo Ball, Sterling Brown, Joe Ingles, Marcus Morris	59	$53,492,617
Last	Khem Birch, Amir Coffey, Jerami Grant, Donovan Mitchell, Carmelo Anthony	25	$57,248,729

Figure 4. We defined the parameters as shown in the first table prior to predicting the wins for each team. Then we obtain the teams that were predicted to have the highest number of wins and obtain the average salary for the teams.

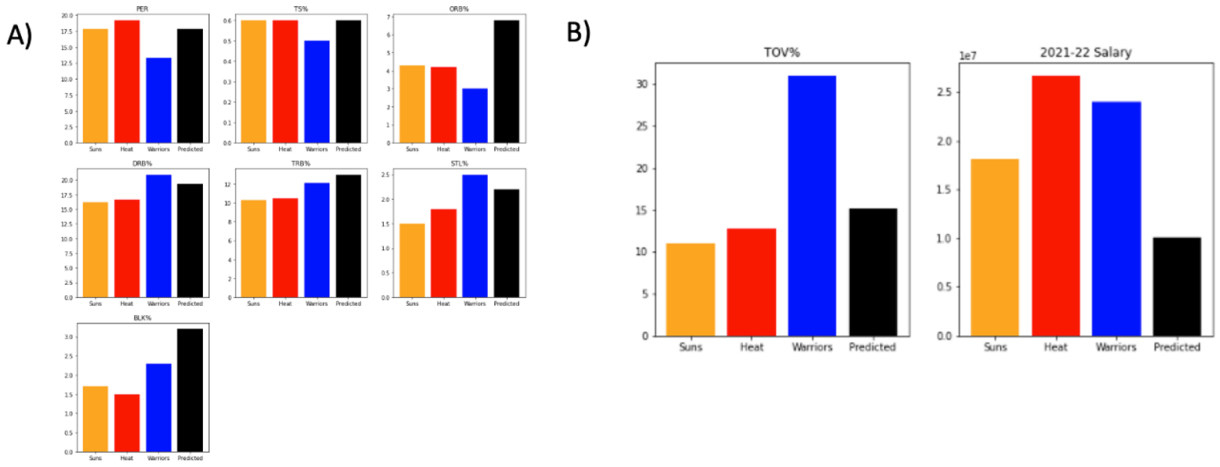

Figure 5. The averaged advanced statistics and salary between current top teams in the 2021-22 season and predicted rank 1 team were compared. Chart A) shows stats where having higher values is better whereas chart B) has stats where lower values are better.

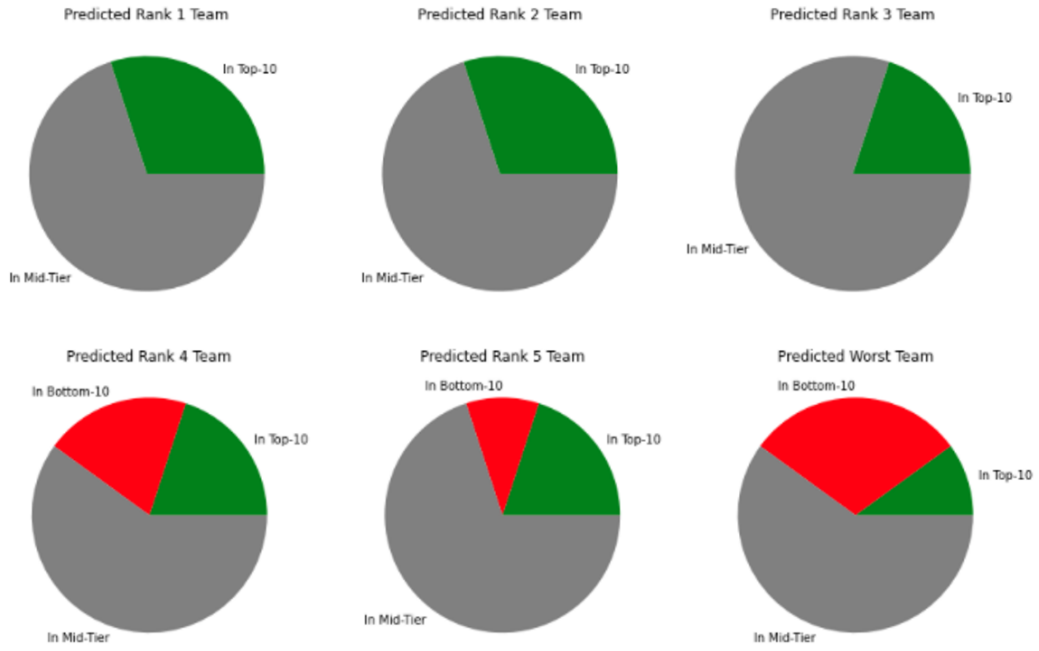

Figure 6. The percentage of players in the predicted teams that belong in current top-10 or bottom-10 teams is shown. The rank 1 predicted team has a greater portion of top-10 teams. More bottom-10 players show up as the team is predicted to have a fewer number of wins.

6. Discussion

6.1 Player Clustering

Our clustering analysis shows five distinct groups of players within the NBA, with each cluster having unique traits (Figure 1). For example, high-performance athletes (group C1 and C2) contribute more to a team's success (based on PER and FIC) and are most likely to be All-Star players. We also see a group of players who are significant assets for a team but do not perform as well as the previously mentioned group (groups C3 and C4) . This group may consist of players a team would add to build a roster around an All-Star player and may have critical roles on a team (i.e., elite defensive/offensive player). Finally, we see a group of players (group C5) that contributes the least to a team's success and do not play many minutes, suggesting that this group consists of bench players.

Comparing the advanced statistics further confirms the differences between each cluster of players and provides insights to the types of players within each group (Figure 2). As previously mentioned, we see groups C1 and C2 both have a high number of All-Star players. However, statically these groups are different. Group C1 appears to be offensive threats, which is suggested by the high number of plays they are involved in (USG%), their offensive impact of the game (PER) and their ability to create offensive opportunities (FIC). Comparatively, group C2, tend to lead in every rebounding statistic (ORB%, DRB%, and TRB%) and have higher blocking percentage (BLK%), which suggest that this group of All-Stars are a “big men” (aka players that use their height and weight to their advantage). Furthermore, these players are effective scores (eFG%), which could be because these players are closer to the rim, and therefore can take easier shots - especially after a rebound. Because group C3 is second in rebounding and blockage, these players appear to be “big men” as well, but don’t play at an All-Star level. Group C4 has the the same amount of playing time (MPG) as the previously mentioned All-Star groups, but aren’t involved in as many players (USG%) and typically contribute to some offensive success of a team (FIC and PER), but not as much groups C1 and C2, which suggest these players are typically starters, but not All-Star players. Finally, group C5 have lowest statistics in every category previously mentioned, which further suggests that these players are typically bench players.

A more granular view looking at the top 20 players (based on PER) further confirms the types of players found in each cluster (Figure 3). All the players in group C1 are All-Star players, and many are considered future hall of famers (Lebron James, Steph Curry, Kevin Durant, Tim Duncan etc.). Group C2, are some of the top “big men” in the league. However this group is small (10 total players), which makes sense since the league has moved away from the traditional “big man” in recent years [7]. Looking at group C3’s and C4’s top players further confirms the initial assessment of players from C3 being a subgroup of “big men” in the league and group C4 being key offensive starters (many having all star appearances or winning offensive awards).

Group 5 also confirms the initial assessment, with many players from this league getting little playing time and not contributing much to a team's success.

6.2 Team Prediction

Figure 4 clearly shows that the predicted rank 1 team has either similar or better averaged statistics when compared against the current top teams in the league. In summary, the predicted team should be capable of grabbing offensive and defensive rebounds, stealing the ball, and blocking the ball at a high percentage. The players would also find success offensively since they have high shooting efficiency. Furthermore, the predicted team would perform well offensively and defensively while also not turning over the ball frequently (losing the ball before making a shot attempt). These players would be performing at this level well under the salary of the top teams as well.

We can also look at the percentage of players in the predicted teams that belong in current top-10 or bottom-10 teams of the 2021-22 season. For example, the rank 1 predicted team has 3 players from a top 10 team in the league and the other 2 are from middle of the pack teams. Figure 5 indicates this breakdown for the rank 1-5 predicted teams and the team predicted to have the least number of wins. There is a clear trend with more players in top 10 teams in the rank 1 team vs. the worst team. We would expect this distribution to become more clear if we increased the team sizes, but we are limited by the historical data of current teams and the number of wins they produced.

7. Conclusion

The hierarchical clustering showed there are five distinct clusters of players in the NBA. Further analysis, comparing advanced statistics, showing differences between groups, confirms a successful clustering model. Finally, investing top players of each cluster, based on PER, provided a final confirmation of the types of players that are found in each cluster.

Our results in the hierarchical clustering allowed us to verify, to a certain extent, that a player’s performance is attributed to his advanced statistics. We use these findings to help support our assumption that we can find viable teams given advanced statistics and historical win data. These teams are cross validated against current successful teams and have shown to match favorably in terms of advanced statistics while being significantly cheaper.

Despite limiting the scope of our team generation because of available data, it is still evident that certain players are paid much less than other players who have similar advanced statistics. The main reasons for this discrepancy are that too much emphasis is put on standard statistics such as points and assists per game. These statistics are likely inflated because of additional minutes played or more free throw attempts. Additionally, our method does not factor

in crucial intangibles such as leadership and basketball iq, which can significantly increase a player's salary.

8. Future Work

Although results may show that the teams predicted are both cost effective and viable offensively and defensively, without generating larger teams, it is difficult to say how this team would actually perform. Teams must rotate through bench players, who are a significant part of a team's success. We would need more data with the different variations of teams and the number of wins they produced. However, getting this data is not feasible since teams rarely make large roster changes. One solution is to take teams from the past into account, but most players do not spend a significant number of years in the league and are not consistently playing at a certain level. If we wanted a more comprehensive and robust algorithm, we would need to determine how many years of player data from the past we could include.

Another feature that can be taken into consideration to further evaluate the players to form a team is their Elo rating. They are the best method to relativize NBA team strength and performance over many years. To calculate elo rating, all teams start at a median score of 1500 and they are either given or subtracted points based on their final score of each game, and where and when it was played. Furthermore, the dataset can be updated with every season which will help with team formation for every year.

As more data is being collected for better prediction and the datasets get larger, dealing with dimensionality reduction with PCA would become difficult. In this situation, PCA can be swapped to PCA + LDA(Linear Discriminant Analysis) which performs better for larger datasets while at the same time preserving much of the class discrimination information as possible. We can also explore other models and see just how they are different from the ones we chose to execute. Some of those candidates could be Bayesian Regression, Ridge Regression, etc instead of using simple linear regression.

References

[1] Lewis, Micheal. 2004. Moneyball. WW Norton.

[2] Triady, Mochamad & Utami, Ami. (2015). Analysis of Decision Making Process in Moneyball: The Art of Winning an Unfair Game. The Winners. 16. 57. 10.21512/tw.v16i1.1555.

[3] Thenmozhi, D. & Palaniappan, Mirualini & Sakthi, S.M.Jai & Vasudevan, Srivatsan & Kannan, V & Sadiq, S. (2019). MoneyBall - Data Mining on Cricket Dataset. 1-5. 10.1109/ICCIDS.2019.8862065.

[4] D'Urso, P., De Giovanni, L. & Vitale, V. A robust method for clustering football players with mixed attributes. *Ann Oper Res* (2022). https://doi.org/10.1007/s10479-022-04558-x

[5] Pappalardo, Luca & Cintia, Paolo & Ferragina, Paolo & Massucco, Emanuele & Pedreschi, Dino & Giannotti, Fosca. (2019). PlayeRank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach. ACM Transactions on Intelligent Systems and Technology. 10. 1-27. 10.1145/3343172.

[6] Nguyen Hoang Nguyen, Duy Thien An Nguyen, Bingkun Ma & Jiang Hu (2022) The application of machine learning and deep learning in sport: predicting NBA players' performance and popularity, Journal of Information and Telecommunication, 6:2, 217-235, DOI: 10.1080/24751839.2021.1977066

[7] Kier, S. (2017, September 28). *The death of the traditional big man in the evolving NBA landscape*. Bleacher Report. Retrieved August 8, 2022, from https://bleacherreport.com/articles/1200804-the-evolving-nba-landscape-and-the-death-of-the-traditional-big-man

Annex-A – This Annex has information about the contribution of each individual in the project group.

Calvin Yu: I worked primarily on the team prediction part including aggregating the data, parsing it, and predicting team wins. I also wrote the abstract, introduction, and parts of the methodology, results, discussion, and conclusion. Lastly, I created my part in the presentation slides.

Daniel Lisko: I worked primarily on the hierarchical clustering section of the project - including web scraping, aggregating and data exploration, data transformation (PCA) and analysis of advanced stats on the clusters that were generated. I also wrote parts of the methodology, results, discussion and conclusion. Finally, I created slides for my part in the presentation.

Ruchi Bhavsar: I worked on the background and future work sections of the report and presentation. I also looked at the correlation of features to help with the feature selection process and did some analysis on the team prediction.

Toan Vang: I worked on the supplemental experiment to predict all-star selections for players and the supplemental experiment portion of the report. I also worked on a few slides of hierarchical clustering.

Supplemental Data and Experiments

Abbrev.	Description	Abbrev.	Description

eFG%	A measurement of efficiency as a shooter in all field goal attempts with three-point attempts weighted fairly	PPR	Pure Point Rating =100 x (League Pace / Team Pace) x ([(Assists x 2/3) - Turnovers] / Minutes)
Total S%	The sum of a player's field goal, free throw and 3pt %.	PPS	Points scored per field goal attempt.
ORB%	A measurement of the percentage of offensive rebounds a player secures that are available to his team.	ORtg	The number of points a player produces per 100 possessions.
DRB%	A measurement of the percentage of defensive rebounds a player secures that are available to his team.	DRtg	The number of points a player allows per 100 possessions.
TRB%	A measurement of the percentage of both offensive and defensive rebounds a player secures that are available to his team.	eDiff	The difference between a team or player's ORtg and DRtg.
AST%	A measurement of the percentage of assists a player records in relation to the team's overall total while he is in the game.	FIC	A formula to encompass all aspects of the box score into a single statistic.
TOV%	A measurement of the percentage of turnovers a player records in relation to the team's overall total while he is in the game.	PER	Player Efficiency Rating is the overall rating of a player's per-minute statistical production.
STL%	A measurement of the percentage of steals a player records in relation to the team's overall total while he is in the game.	GP	Number of games played.
BLK%	A measurement of the percentage of blocks a player records in relation to the opponents two point field goal attempts.	MPG	Average number of minutes a player has played per game.
USG%	A measurement of the percentage of plays utilized by a player while he is in the game.		

Table S1. Abbreviations (Abbrev.) for all the advanced statistics used in this study and their description [7].

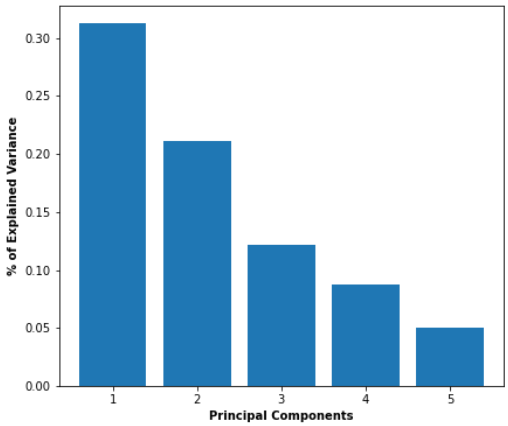

Figure S1. Percent explained variance for the top 5 components generated from a principal components analysis on advanced NBA statistics.

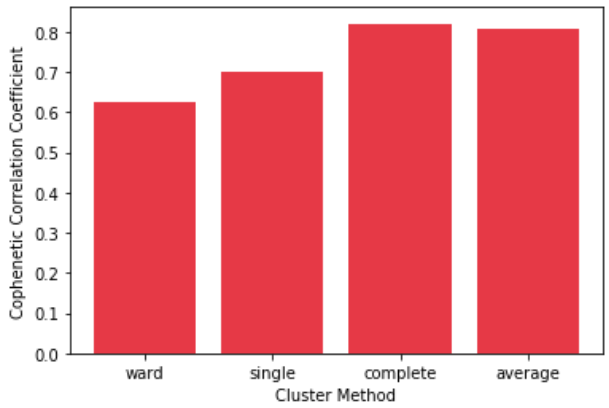

Figure S2. Cophenetic correlation coefficients for Ward's method, single linkage, complete linkage and simple average clustering methods.

Supplemental Experiment 1: Predicting the Number of All Stars Selections Per Player

A decision tree (DT) and k-nearest neighbor (KNN) model were used in order to predict the number of all-star selections per player. Both of them provided high accuracy but the DT performed better than KNN. The DT had an accuracy of more than 95% for the test set, whereas the KNN only had an accuracy of 83% for the test set.

A majority of the players only received 0 or 1 selections. As a result, the data points are not evenly spread. Since KNN relies on neighbor values and distinct groupings, it struggled to predict accurately. For example, only one player has 14 all star selections and the same with 13 selections so the KNN groups all these people as one group. On the other hand, the DT model groups data based on the frequency range of data points which increases the chance of the correct prediction. If the data points are interrelated with each other then the DT will group based on the distribution of the data points.

0	745
1	37
2	19
3	13
6	9
4	8
5	8
7	6
8	5
10	5
15	3
12	2
18	2
9	1
11	1
13	1
14	1

Figure S3. Frequency of All-Star Selections

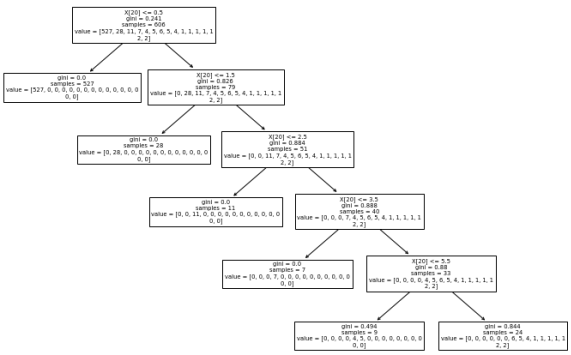

Figure S4. Decision Tree Splits