Can Stats Alone be Used to Predict a Champion?

Matthew Gerola

CptS 315: Spring 2022

Course Project Report

Introduction:

The data mining task that I have chosen is to try and predict the champion for the current and future NBA seasons based on regular season stats. The motivation behind this project stems from the importance stats now play in sports, with stats dominating decision making. Whether to take a three in basketball, go for it on fourth down in football, or shift the defense in baseball, all these decisions are now based on stats from previous outcomes and the possibility of success. Which is why I want to see just how powerful stats are by using them to try and predict a champion.

Some questions I will be investigating in this project include whether stats alone can be used to predict the champion for a season accurately as well as how reliable is the centered cosine similarity to assist in the process of predicting a champion for a sports season.

The personal motivation behind choosing this as my project stems from my enjoyment of sports and the importance stats are now playing in sports, with more advanced stats being taken each year. My goal for this project is being able to use 21 years of data to accurately predict the champion for the current and future NBA seasons using only stats.

Some of the challenges I face when doing this project are how to filter the teams to only have one team be left at the end, along with how to calculate stats for the following year from the data already provided. In addition to those challenges I need to know how confident I am in the predictions I made, which is another challenge I face. My approach to the task involves taking the similarity each team has with past champions and having the top three highest similarities in each conference go into a simulated playoff series where a champion will then be predicted. To predict the stats for the following year I will get the stats from the past three seasons and have a weighted average to get the stats for the following season and then predict a champion from there.

The results I got from this project are that the centered cosine similarity is a useful method to help filter out teams, but that stats alone may not be enough to accurately predict the champion for a given year. This stems from the reality that stats change when playing in the playoffs either for the better or worse, which would make the regular season stats not as valid to predict the champion.

Data Mining Task:

Get the regular season stats for each team from the past 21 seasons with the champions known via the official NBA website, plus the regular seasons stats for this year's teams as well. Use the stats from the past champions to get the six most similar teams for the current year, three from each conference by implementing the centered cosine similarity. From those six teams simulate a playoff series to predict the champion for that year. After this task is done, predict the stats for the following season based on stats from the previous three seasons and do the same task, get the six teams most similar to the past champions using the centered cosine similarity, predict the champion from those six teams through a playoff series and arrive at a predicted champion. With the output being for both the current and future seasons the final six teams and the predicted champion. With a percentage to indicate how confident the future champion is present in the final six and how confident the predicted champion is going to be the champion for the given year.

Example of the Input Data:
Golden State Warriors
79,63,16,.797,48.1,115.0,43.7,89.7,48.7,11.6,31.7,36.5,16.1,20.2,79.8,9.5,35.6,45.1,27.4,13.0,8.5,4.4,3.9,19.7,19.0,7.8

Example of the Output Data:
"Final Teams 2021-2022: Phoenix Suns, Milwaukee Bucks, Brooklyn Nets, Golden State Warriors, Miami Heat, Utah Jazz: 90%
Predicted Champion 2021-2022: Milwaukee Bucks 75.7%"

Questions Looking to Investigate:

1. How accurate is the centered cosine similarity when it comes to sports, i.e. how reliable is it to help predict the champion for a season?
2. Can stats alone be used to predict the champion for a given year and following year?

Challenges Faced to Solve this Task:

1. How to simulate a playoff series to determine the champion with the final six teams after the filtering process.
2. How to filter out teams for the playoff simulation.
3. How to determine the stats for the following year.
4. How to get a confidence number when making the predictions for future champions.

Technical Approach:

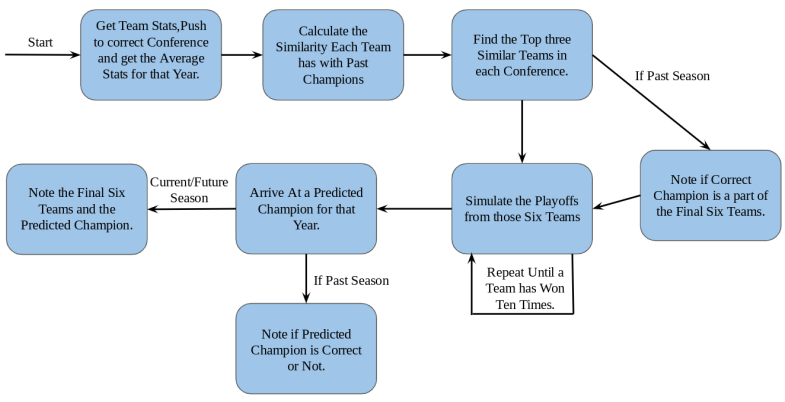
To begin the task I would get the stats from the 2000-2001 season. For this season I would only be taking note of the stats for the champion and the averages for that season to normalize the champion's stats. I would then push the champion with their normalized stats to a vector to be able to compute the similarity with teams from other seasons. After this I would get the stats for the 2001-2002 season, but for this season I would get both the average stats and each team's stats with each team being pushed to its corresponding conference vector. I would get the conference for each team from another vector I had populated prior and traverse until the team was found with the conference found tied to that team being the vector the team was pushed to. After the teams and stats have been read in I would call the function to find the total similarity each team has with the past champions in this case the 2000-2001 champion would be the only past champion used. I would find how similar each team's stats are with the past champion's stats by implementing the centered cosine similarity. After all the teams in each conference have a similarity value, I would traverse the two conferences and find the top three highest similarity values in each. This is the process I will do to solve the challenge of filtering out teams for the playoff simulation. After the top three most similar teams in each conference are found, they are noted and the playoff simulation function is called. The way I solved the challenge of simulating a playoff series with the final six teams is to first have the team with the highest similarity get a "bye" and not play in the first round with the other two teams in the conference "playing" each other, with the simulation of a playoff series specifically a game being as follows. I would pass into a function the total amount of two pointers attempted which is the total field goals attempted minus the three point field goals, the amount of three point field goals attempted and the amount of free-throws attempted, together with the percentages made for each shot. I would have a loop that is equal to the amount that shot is taken, with a random number called for each iteration, and if the random number is less than or equal to the percent that team makes for that shot then that shot's point value gets added to the teams total for that game. The winner would be declared as

the team with the higher amount of points for that game. In the case of tie breakers it would go as follows: if the points are equal, the team with the higher winning percentage for that season would be declared the winner, if those values are equal then the next tiebreaker is the team that averaged more points that season. This process of simulated games would go on until one team has won four games since an NBA series is best of 7. This would be the same process for the proceeding series as well, the only difference being the teams playing, with the winning teams moving on and the losing teams not. After all the series are concluded a champion is crowned, I made it so a team would need to win the championship a total of ten times before being declared the "predicted" champion. I made this decision to try and be more certain about who would actually win, with the rationale being if they won ten times then it must not be a fluke. I would do this for all the past seasons, and with each increase of season the correct champion's normalized stats would be added to the past champions vector, so they can then be used when calculating the total similarity teams have with past champions in later seasons. This means when the current season is looked at a total of 21 past champions will be used in the similarity process. A step I would do with the past seasons only is to note whether the actual champion is present in the final six teams and whether the predicted champion is correct. If the actual champion was present in the final six for that year, a one would be placed in an array or zero if not present. This would be the same for the predicted champion as well if the prediction was correct a one would be placed in an array otherwise a zero. So when all the past seasons are finished the totals in each array would be added up and divided over 20 with a percentage to indicate how confident the future champion is a part of the final six, as well as how confident the predicted champion will win for both the current and future season. This is how I found a confidence level and solved that challenge for this task.

After all the past seasons were run on the algorithm and a confidence found for the predictions to follow, I would move onto the current year's stats. I would do the same procedure as above for the current year. I would get all the teams with their stats, push them to their correct conference and get the average stats as well for that season to normalize the teams. I would calculate the total similarity each team has with the past championships, 21 past champions to be specific, with the centered cosine similarity and have the top three in each conference move onto the playoff simulation. I would note the final six teams to be printed out later, run the playoff simulation and have a predicted champion that would also be printed out later. After running this part of the task I would move onto the future champion calculations.

To predict the future stats for each team I would first get the stats from the past three seasons for each team. Next I would do a weighted average to have the more recent years have a greater impact going into the following season. Which is how I solved the challenge of getting the future stats, by going with the weighted average approach. I would then do the same thing as before, populate the two conferences with the correct teams and their predicted stats, calculate the total similarity with the past champions including the champion I predicted for the current year, get the top three teams from each conference, simulate a playoff series from the top three teams in each conference and arrive at a predicted champion with the final six teams and the predicted champion being printed out.

Diagram of the Process:



Evaluation Methodology:

The dataset I used in this task comes from the official NBA website, under the regular season team stats section. With the basic stats collected during an NBA season as in wins, losses, win percentage, average points scored in a game being used in the task. This makes it an ideal dataset to use since it's from the official site of the NBA meaning it will have all the official/correct records, which is important for this project since it is exploring stats and the power they hold for predicting future events. The data was fairly easy to use for the task I set out to do. The only real challenge this dataset had was the way stats were separated, that being with spaces specifically tabs. To combat this I decided to format the data with commas with the team beforehand then on the next line the stats they obtained for the given year. This wasn't necessary to be able to use the data, it was more to have the program have an easier time traversing the data for each year. The way I would evaluate the output of the program for the current and future seasons since they are the ones being predicted, comes from the program being run on the past seasons and the confidences found from doing those seasons, which would end up being percentages. I would then use those percentages to evaluate how well this iteration of predictions were. The percentages would also help to answer the questions of "How accurate is the centered cosine similarity when it comes to sports, i.e. how reliable is it to help predict the champion for a season?" since it was used to filter out teams and "Can stats alone be used to predict the champion for a given year and following year?" since stats were used in both the filtering process and playoff simulation. This would match real-world applications because I would be testing the algorithm on past events with results already known to then know whether the algorithm and the approach was producing results that were accurate.

Results and Discussion:

The results I got from this project were interesting, both in the filtering out of teams and in the simulated playoffs. The result from the filtering out process showed that future champions should be fairly similar in stats to past champions. This conclusion comes from all the champions of the past being a part of the final six teams for their particular season, which were found from the centered cosine similarity. This means the centered cosine similarity is a good mathematical equation to employ when dealing with stats, with the goal of filtering out the best teams, but not necessarily the champion from it, which was the goal. Which goes into why I chose to include three teams from each conference instead of only two. With only the top two teams from each conference being used the correct champion was a part of the final teams in only 16 of the 20 possible seasons which is equivalent to only 80%. After I increased the teams to three per conference the amount the final teams contained the correct champion increased to 20 out of 20 seasons making it 100%. An example of this happening can be seen on Figure 1, with the Milwaukee Bucks being left out if only the top two teams from each conference were used even though they would eventually be the champions.

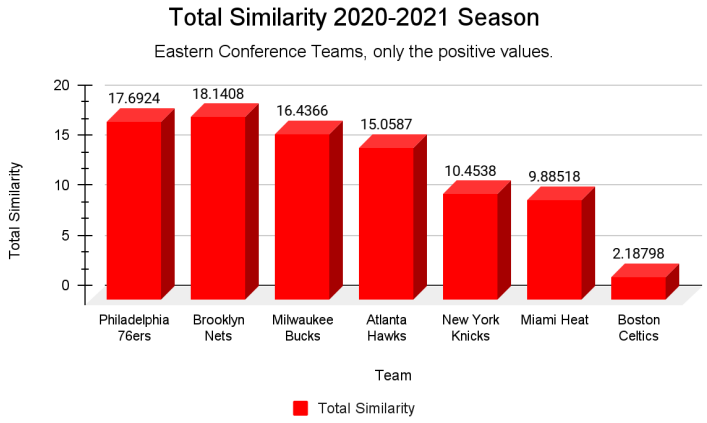


Figure 1- The total similarity a team shared with past champions

Following the filtering of teams was the playoff simulation, these results surprised me, because the number of correctly predicted champions from the past was lower than I expected. The percentage increased even if slightly though, when I added an additional team from each conference. This was initially to get all the previous champions into the playoff simulation from the filtering process, but it also helped to slightly increase the predicted champion percentage, so in the long run that was a good change in the algorithm, this can be visualized in Figure 2. I initially thought the issue of having a low number of correctly predicted champions was due to the points being generated in a game for a team being incorrect, which would result in the wrong team moving forward in the playoffs and hence an incorrect champion in the end. This though was not the case as outlined in Figure 3. The points were fairly close to the team's average for the year, but there are some noticeable outliers. The outlier though could be explained as a team either having an exceedingly good or bad game. So that is one place in the algorithm where an issue wasn't present. So that led me to thinking about what was going on during some years where the predictions were wrong, and if the same players in the regular season were present during the entire playoff run that the team went on. Meaning if a team was missing one or multiple key players during the playoffs the stats would be noticeably different than the stats generated in the regular season and vice versa if a team was missing one or multiple key players during the regular season and they came back in the playoffs the stats would be different as well. Which is why I think the predicted champions are different from the actual champions, the team's stats changed when the playoffs happened.

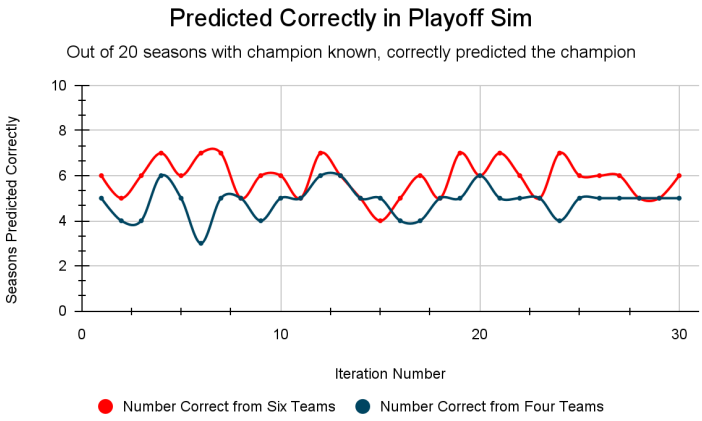


Figure 2- The amount of correctly predicted seasons

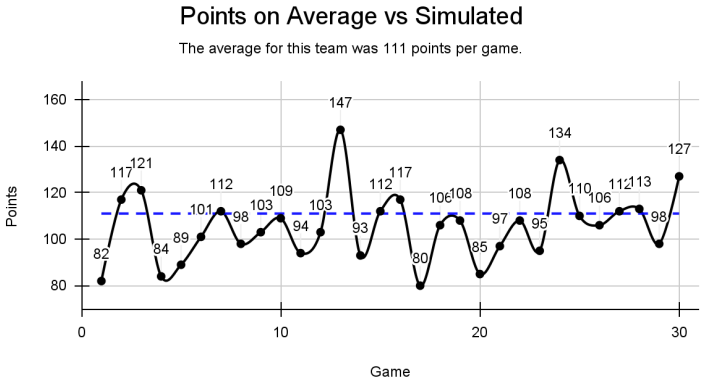


Figure 3- The amount of points on average per game versus simulated for a game.

So a part of my approach to this task that worked was using the centered cosine similarity to assist in getting three teams from each conference. I know this approach worked because in all the past seasons the champion was a part of the final six teams that would move onto the playoff simulation. Since this approach worked it would follow that future champions should be fairly similar to past champions, hence why it succeeded in the end. This leads into a part that didn't work as well though, the playoff simulation. In the playoff simulation it assumes each team is going to be able to put up the same stats in the playoffs as they did in the regular season. So the simulation would favor the team that has the best stats in the regular season, but in the playoffs if

a team gets hot and if they are good enough they could still win the championship. Which was evident in the filtering process: all the champions were a part of the final six, but some weren't a part of the final four, which was the initial number of teams to include for the playoff simulation. This means a team can have a high similarity to the past champions, but not necessarily the highest, maybe not even the top two but still win the championship. So if they weren't statistically the best of the final teams or the best in their conference they wouldn't be picked to win the championship therefore an incorrect prediction would be made for that year. Which is why I think that this part of the approach wasn't successful, which was further supported with the low percentage of correctly predicted teams.

From the results gathered I can answer both questions that were asked at the beginning with some sense of confidence. The first being "How accurate is the centered cosine similarity when it comes to sports, i.e. how reliable is it to help predict the champion for a season?". The answer I have come to with this question is that the centered cosine similarity is extremely valuable to assist in the process of choosing a champion. I say assist because from the data gathered choosing the champion based solely on similarity is not a great idea because the champion could be the third highest similarity value in a conference to the past champions which was proven from the inclusion of one extra team into the final teams since all the champions were then present afterwards. The question of "Can stats alone be used to predict the champion for a given year and following year?" though is a different answer. Stats can be used to a degree to help predict the champion, but not with a great deal of confidence. This comes from the fact that a player could get hurt or a player could come back and change the stats the team produces during the playoffs. This would result in a team not producing the same stats, which could be in the team's benefit or detriment and result in a team that wasn't necessarily the best winning the championship or the best team not winning the championship.

Lessons Learned:

Some of the lessons learned from doing this project are the following, future champions are fairly similar to past champions. Along with the fact that the team that is the best statistically in the regular season or the most similar to past champions doesn't necessarily mean they will win the championship for that year since things can change in the playoffs either for the better or worse.

Some things that I would do differently if I could do the project over again would be to include the playoff stats as well as the regular season stats. So when the final six teams were found the playoff stats would be used in the playoff simulation instead of their regular season stats. I would do this in the hope that the correctly predicted number of champions would go up when the program is run over the past 20 seasons.

Acknowledgments:

NBA website where the stats were obtained from:
https://www.nba.com/stats/teams/traditional/?sort=W_PCT&dir=-1&Season=2021-22&SeasonType=Regular%20Season