Akhil Chamarthy
Pranav Jayaraman
Anirudh Vadrevu
Henry Vuong

Predicting NBA Regular Season Records to Find Undervalued or Overvalued Playoff Teams

Problem Statement

The National Basketball Association (NBA) playoffs are the pinnacle of professional basketball - it is where legacies are defined and champions are crowned. An estimated 20 million Americans[1] tuned in to see the 2018 NBA Finals, and considering the popularity of the NBA internationally, millions more watched the Golden State Warriors claim their second title in a row, and third in four years. The NBA Playoffs are also a ripe opportunity for bettors and casinos to cash in. This year's playoffs are the first since sports betting was legalized in New Jersey, Mississippi, West Virginia, New Mexico, Pennsylvania, and Rhode Island[2], and considering an estimated $4.5 billion[3] was spent on betting in Nevada in 2018, millions (potentially billions) will be bet on the playoffs this year.

The NBA Playoffs are structured as followed: the top 8 teams in each of the 15-team conferences (the East and the West) play against one another based on their seeding, where the 1st seed team had the best regular season record and plays against the 8th seeded team in that conference. They play one another in a seven game series in each round, and the overall winners from each conference meet in the NBA Finals. Vegas and sportsbooks place a premium on the teams with higher seeds since they had earned the better records throughout the regular season, will have home-court advantage, and are seemingly "better" than lower-seeded teams. However, an inherent problem from this method is that some teams are undervalued by Vegas, as they should have had a better record but did not; and vice-versa, some teams are overvalued.

We took the problem of determining how to properly evaluate NBA playoff teams and decided to build models that used regular season statistics, along with other important features, to attempt to predict what an NBA team's record should have been. By building such a model, bettors can bet on undervalued teams and against overvalued teams. Furthermore, this same approach can be used by NBA teams themselves as an analytical tool to predict their regular season records and make appropriate roster changes in advance. It is beneficial for a team to finish with a great record to have a higher seed because it results in home-court advantage, which is generally extremely beneficial. Teams can use their current roster to extrapolate statistics for the next season by using this model, and they can predict what their record would be given their current roster and which statistics make the biggest impact on the regular season record. Teams could then pursue free agents or make trades for stars that will improve their team.

Solution

To analyze this problem, we decided to use regular season NBA data from the past 10 years. We scraped this data from the official NBA website, and added several of our own columns as well. The

[1] https://deadline.com/2018/06/kevin-durant-warriors-win-nba-finals-ratings-down-nba-abc-1202406923/
[2] http://www.espn.com/chalk/story/_/id/19740480/gambling-sports-betting-bill-tracker-all-50-states
[3] http://www.espn.com/nba/story/_/id/21597720/nba-how-nba-preparing-cash-legalized-sports-betting

official NBA data had dozens of categories, including Traditional, Advanced, Four Factors, Misc, Scoring, Opponent, and Defense; each of these categories contained different pieces of data. For example, true shooting percentage (a statistic that measures how efficiently a player shot the ball) fell under Advanced, but field goal percentage (the number of shots made over the number of shots taken) fell under Traditional. In an effort to further understand basketball statistics, our team read the "Basketball on Paper: Rules and Tools for Performance Analysis" by Dean Oliver. Although the book was interesting and provided us with insights on the meaning behind specific statistics, we decided to use every statistic found on the NBA website to make our model as accurate as possible. Furthermore, we created some of our own features, such as the Number of All-Stars on each team (the All-Star game is a mid-season exhibition game played between the top 20 players in the league, selected by the fans, media, and coaches), and Number of All-NBA First, Second, and Third Teams (an NBA end-of-season honor comprising of the top 15 players in the NBA).

Before analyzing the data, we combed through the data and found some interesting results. When comparing defensive rating against win percentage, we found that a lower defensive rating correlated to a higher win percentage (see Appendix A). Exploring the data gave our team a better understanding of some of the analytical statistics that the NBA uses in today's day and age, and disproved some of our long-held assumptions, such as the importance of defense for winning games. After spending a significant amount of time gathering and cleaning our data, we began to analyze ways to feature engineer our data. We decided to bin our classification label, which was winning %, into 5 different bins, each with a different range of winning %. This allowed us to use winning % as our label because Darwin did not allow for labels to be continuous variables.

The Darwin supervised learning module was used for our data analysis. Initially, we had assumed that Darwin would create a model that had feature scaling performed. However, our Darwin generated model continued to use all 92 features and produced subpar accuracy. To increase accuracy and achieve feature scaling, we selected only the top nine features and Number of All-Stars and ran Darwin against only these features. The accuracy of our model increased from 61% to 80% accuracy. This is an area where Darwin should understand on its own that the other 80 features were poorly correlated and exclude them from the prediction model. Another feature that Darwin was missing was cross-validation when splitting the dataset into testing and training. Darwin requires the users to split the data set before uploading to Darwin when it would have been better to have an option for cross-validation. This would prevent the training set from possibly missing patterns found only on the testing set and save time from having to manually split the datasets. Darwin also lacks was support for continuous labels. Having continuous labels would broaden the application of Darwin and eliminate some of the shortcomings of binning. Selecting the number of bins and where to bin is about balancing accuracy and information loss and having continuous variables would have lessened the challenge of binning.

As stated earlier, our final model could be used to find under and overvalued playoff teams. Given our model's accuracy (80%), we are confident that it can be used to predict over and undervalued teams in this year's playoffs. If a team is overvalued, sportsbooks likely place a premium on that team, so betting against the team could lead to increasing winnings. The same could be said about the opposite; betting on an undervalued team, which likely does not have a premium placed on it, could significantly increase earnings. Doing such an analysis could revolutionize the way average NBA fans approach and bet on playoffs. But this type of analysis can also be done by NBA teams looking to make their teams better. An NBA team could gather more granular data and using their expertise could select features that could make the model more accurate.

Team Engagement

The team divided roles by having Henry and Akhil working together with the Darwin system to analyze our data, and Anirudh and Pranav doing background research, collecting the raw data, and

cleaning it in order to be properly processed. We were all part of brainstorming ideas portion, and we all talked through roadblocks when collecting & cleaning data and using the Darwin system to obtain information. Team participation overall was good with Henry and Anirudh taking the initiative to ensure that the Darwin system was compatible with our needs.

General Challenges

The first challenge we faced involved our topic selection. Our initial idea focused on combine data for incoming NFL rookies, but the data we had collected had too much inconsistency and missing information, which caused us to pivot our focus to our current topic. The data collection phase was a difficult process as some of the information we needed was not readily available though, such as the number of all stars on a team every season, and we had to search through years of stats for each team every season in order to obtain this information. However, the data collection and cleaning portion of the assignment overall worked well. Darwin was a more difficult to work with than anticipated because it would reject our requests to train model sometimes, which would create challenges as well (probably as a result of too many requests to the server). There seemed to be many limitations in Darwin's system that caused us to have to pivot and shift our focus for the project because the initial goals for what we were solving changed when we realized some of the functions that Darwin is unable to support. Some challenges that we faced with Darwin included Darwin's inability to properly exclude poorly correlated features, lack of a cross-validation checking system, and failure to use continuous labels, which could potentially provide more insightful information. We had tried to create a clutch metric as another unique statistic that we feed in with the rest of our data, but this did not work successfully because there were issues with creating the measurement for it.

Next Steps

Overall, the next steps that would be required to take a deeper dive into this project would be to collect even more data from older seasons of basketball when the game was played differently than today's generation, and see if the same factors that correlate to accuracy with our current data still hold true or not. If so, this would be a signal to betters and NBA teams that certain statistics can stand the test of time no matter the era and play style of basketball, which can heavily influence the future of NBA analytics and future growth. Additionally, it would be beneficial to use a continuous model in order to have more valuable information that could be more accurate than binning. Lastly, another attempt could be made at creating a clutch metric for each NBA team, which would better legitimize close wins for certain teams.

Running our model through the 2019 NBA season produces a 78% weighted accuracy. There were a total of eight teams total that over/under performed expectation, four of those teams being playoff contenders. The Denver Nuggets, LA Clippers, and Houston Rockets won more games than expected and the Utah Jazz won less games than expected. Therefore, one might assume that the series between the Rockets and Jazz (currently going on at the time of this paper, with the Rockets leading 3-0) would be closer than expected. However, the Rockets lead the Jazz 3-0 in the series. Although the Rockets won more games than expected, they are still a very good team (this roster did make the Western Conference Finals last year and take the champion Warriors to 7 games), and so maybe an undervalued Rockets team is still an incredibly talented roster. The Denver Nuggets are currently playing the San Antonio Spurs, and as a 2-seed the Nuggets are heavily favored. But our model shows the Nuggets are overvalued, and the series is currently tied at 2 games each. Our predictions are currently a mixed bag, with some being more accurate than others, but given more data, we are sure our model will be more accurate.

Appendix

Appendix A