Predicting the Results of NBA Games Using Machine Learning

Sam Macpherson, Konstantin Topaloglou-Mundy, Lazar Vukoje
{skmacphe, ketopalo, lvukoje}@uwaterloo.ca
University of Waterloo
Waterloo, ON, Canada

Introduction

In recent years, the NBA has been extending its reach and attracting an increasingly larger and diverse audience. NBA games are viewable in many nations across the globe, and in the 2014-2015 season, the league featured 92 international players, hailing from 39 different countries (Jones 2016). In 2019, an average of 15 million people—per finals game—watched Kawhi Leonard and the Toronto Raptors garner their first NBA championship (Gough 2019). The American-based league, at the age of 74, now sits in fourth place among all sports leagues in the world by revenue (Amoros 2016). People from all over the world enjoy watching and cheering on their favourite teams, and as the NBA increases its global presence, we see an equal increase in the revenue the league pulls, not just from broadcasts, but from fans purchasing jerseys and other merchandise as well. Over the last few years, we have seen a rise in a separate—though related—industry: sports betting. On May 14, 2018, the federal ban that was in place in the United States which prohibited betting on the outcomes of sporting events was lifted, leaving states free to decide to pass legislation to legalize sports betting (Licata 2019). Since the ruling, 11 states have legalized sports betting—amoung them Nevada, Oregon, New York—and an additional 24 states are in the process of passing legislation (Licata 2019). With the recent legalization efforts, the derivative industry has witnessed enormous support: over $21 billion in total US handle (amount of money wagered) since 2018 (Legal Sports Report 2020). It goes without saying, then, that the ability to accurately predict sports results—especially with a mainstream league like the NBA—can be highly lucrative. This paper will describe our efforts to use existing game and player data in order to predict the result of future, arbitrary NBA games. We will focus on training a machine learning model using a season's worth of data as the input features. With high variance in team composition and individual player ability within the NBA from year to year, we believe it prudent to limit the input features to a single season, and use a model trained by such features to predict games from the playoffs for that season. We will take into account team statistics such as points per game, field goal percentage and turnovers. It is important to note that many other—sometimes abstract—outside factors can impact the result of a game, such as the "home court advantage". To limit the scope of this project, we will limit our analysis to concrete statistics.

Related Work

The selection of variables that contribute to a basketball game has been a primary element of study in related work on this topic. The research conducted in "Predicting Outcomes of NBA Basketball Games" (Jones) used over 20 different game statistics as features for their models, as well as four different approaches: 3-game moving average of team stats, 3-game moving median, 3-game moving weighted average, and seasonal averages. One technique of particular interest was that the statistics for a given team would not only include the teams statistics but also include the statistics of opposing teams when they played the team in question. Another interesting technique in this paper was the use of preprocessing on statistics; the author attempted to measure the significance of stats in training the models in order to eliminate statistics that were deemed to be unnecessary or insignificant. However, the author used data from the previous 3 seasons to predict NBA outcomes, which may be a faulty technique because of potentially significant changes to team rosters from season to season. In contrast, the authors (Cheng et al. 2016) of the paper "Predicting the Outcome of NBA Playoffs Based on the Maximum Entropy Principle", restricted the data to only the current NBA season. They also used a different technique: a max entropy classifier. They argue that a max entropy classifier is appropriate for predicting NBA data for two reasons: the first is because it does not assume features are conditionally independent (which would be a mistake for NBA statistics), and second, because max entropy classifiers tend to perform well in situations with smaller number of samples (once again the case with NBA statistics). The methodology in their experiment was to use the stats of two NBA teams from the regular season of a given year as features and then label the data with the results of the playoff matchup between those two teams for the same year. This approach seems more logical than what is was presented by Jones, because the personnel of a given team may change from season to season and so data from prior NBA seasons may not be of use for predicting an NBA playoff outcome. For example, it would not make sense to use the Golden State Warriors' dominant 2014-2018 seasons—where they maintained a win percentage of over 70%—to predict the outcome of the 2019-2020 matches, where after sustaining multiple injuries and a change of roster, the Warriors struggled to maintain a win percentage of 23%. In regards to the stats Chang et al. used, they decided to only use 14 rudimentary team stats (points, assists, steals, blocks, field goal percentages, etc.). To further the discussion of relevant game statistics, a third paper, "Identifying Basketball Performance Indicators in Regular Season and Playoff Games" (Garca et al. 2013), found that in regular season NBA games, the most important indicators of team victory was dominance in assists, defensive rebounds, and successful 2 and 3-point field-goals. However, in the NBA playoffs, they found that the only strong indicator was defensive rebounding. The importance of defensive rebounding is corroborated by the work in "Relative importance of performance factors in winning NBA games" (Teramoto and Cross 2010), and the authors offer an important potential explanation for this: in the playoffs, the two competing teams likely have similar shooting efficiency. This insight could be important for our work, as there will likely be matchups where a conceptual "tiebreaker" is needed to increase the accuracy of the prediction, and this stat is evidently a candidate.

Methodology

Firstly, we intend to scrape data from https://www.basketball-reference.com/ in order to get the data we need to train our models which we will accomplish by using/customizing to our needs, an open source API we found online. We will then look into how we can attempt to formulate our data so as to account for the many factors that can affect the winner of a game of NBA basketball. Finally we will look through some pre-implemented ML algorithms (most likely ones provided by Tensorflow as some of the group members have used Tensorflow before) and select 2-4 that we will train using the same data and compare on how well they predict the outcomes. We believe that the implementation of this project will be relatively straight forward.

Results

Based on the papers we have read online, it seems at though none of the models have been able to predict NBA outcomes with an accuracy of much more than 75%. We do not anticipate to be able to predict outcomes with super high accuracy, but what we do hope to discover is if there is a certain type of algorithm that can train a model that is significantly better at predicting NBA outcomes than others based on the same input data.

References

[Amoros 2016] Amoros, R. 2016. Top professional sports leagues by revenue.

[Cheng et al. 2016] Cheng, G.; Zhang, Z.; Kyebambe, M.; and Nasser, K. 2016. Predicting the outcome of nba playoffs based on the maximum entropy principle. *Entropy* 18:450.

[Garca et al. 2013] Garca, J.; Ibez, S. J.; Santos, R. M. D.; Leite, N.; and Sampaio, J. 2013. Identifying basketball performance indicators in regular season and playoff games. *Journal of Human Kinetics* 36(1).

[Gough 2019] Gough, C. 2019. Nba finals average us tv viewership 2002-2019.

[Jones 2016] Jones, E. S. 2016. *Predicting outcomes of NBA basketball games*. Ph.D. Dissertation, North Dakota State University.

[Legal Sports Report 2020] 2020. Sports betting revenue tracker us betting revenue and handle by state.

[Licata 2019] Licata, A. 2019. 42 states have or are moving towards legalizing sports betting here are the states where sports betting is legal.

[Teramoto and Cross 2010] Teramoto, M., and Cross, C. L. 2010. Relative importance of performance factors in winning nba games in regular season versus playoffs. *Journal of Quantitative Analysis in Sports* 6(3).