DataBall
Predicting NBA Winners with Data

Kevin Lane

August 5, 2017

1 Introduction

The world of sports betting can be very lucrative for those that are skilled at it. The sports bettor Haralabos Voulgaris makes a million dollars in a "bad" year [1]. This potential payout coupled with the easy access of online gambling leads sports betting to be a big industry. An all-time high of $4.2 billion was bet on sports in Nevada in 2015 [2]. However, it is difficult to be a consistently successful sports bettor. Oddsmakers are very good at their jobs and teams' records against the spread and over/under lines do not typically deviate far from 50%. A good example of this is the 2015-16 Golden State Warriors, who set an NBA record for wins with 73. According to covers.com, they were a mere 45-35-2 against the spread and their over/under record was nearly identical at 45-36-1 [3], meaning their games hit the over 45 times.

Betting on the spread involves picking winners of games with modified scores where the favorites give up points to their opponents. For example, if a given team is favored by 5 points, they must win by more than 5 points in order to "cover" and win against the spread. The betting line for this scenario will be set at -5 for the favorite. The opposite holds true for underdogs and losing by less than a positive line. The over/under requires comparing the total number of points scored by the two teams in a given game to the line set by the oddsmakers. Bettors must decide if they think the total points scored will be over or under this value. These provide more challenging problems than simply picking game winners. I plan to build models that predict NBA game winners against the spread and whether or not games will hit the over. These models will provide some guidance into which games are more promising to bet on.

2 Background

This project is a continuation of another project where I predicted the winners of NBA games straight up, meaning I predicted winners regardless of the betting lines. That project is not applicable to a betting scenario since I used season-averaged stats and bettors will only have stats available from the games up to the game they are betting on to aid their decisions. However, it did provide some valuable lessons learned that can be applied here, such as what algorithms and features work well. The first project can be found here and the code repository is stored on GitHub here.

I chose to focus on basketball primarily because the NBA is arguably the most deterministic of the major American professional sports leagues, providing an easier problem to tackle compared to other sports. The best teams win more often than in other sports. At least one team wins more than 60 games (winning percentage of 73.2% in an 82 game season) nearly every season. Contrast that with baseball where randomness plays a much bigger role. Teams winning 100 games (winning percentage of 61.7% in a 162 game season) or more does happen, but not nearly as often. The best MLB teams typically win only about 60% of their games in a given season. A similar trend exists for the worst teams. The worst teams in the NBA win about 20% of their games, while the worst in MLB win about 40%. This results in a wider range of team performance in the NBA than in MLB. The NHL is also notoriously random. The best teams in the NFL often win a high percentage of their games, but the sample size in a given season is small at only 16 games. A team that goes on a lucky winning streak can have a record at season's end that is inflated compared to its underlying talent level.

3 Data

Another reason I chose to look at the NBA instead of another sport is the amount of data readily available. The NBA provides a wealth of basic and advanced stats on their website stats.nba.com. The site exposes a wide variety of information in JSON format through several web API endpoints and parameters that take the form:

stats.nba.com/stats/endpoint/?params

making it easy to obtain the data programmatically. For example, individual player stats from every game of the 2016-17 season can be found here. I plan to pull the necessary stats by utilizing an existing GitHub project that provides a simple API and returns queries as Pandas DataFrames. I will combine the stats with spreads and over/under lines obtained from covers.com, which provides information going back to the 1990-91 season. I will utilize the Python web scraping framework Scrapy to parse the raw HTML and pipe it to a SQLite database.

4 Outline

I plan to build classification models to predict NBA game winners straight up and against the spread, as well as predict if games will hit the over based on the strengths of the two teams in question. The model will be trained with games from previous seasons and tested against the 2016-17 season. The difference between this and the first project is that this project will average stats from previous games to predict a given game, whereas the previous project used season-averaged stats.

I will also experiment with how much data to use to predict a given game. The window size (number of games to use) will be a model parameter that must be tuned for maximum performance. I will track the model's accuracy throughout the season. It will likely be inaccurate at the beginning of the season, but will better predict games as the season progresses and more data becomes available. Lastly, I will track return on investment had bettors placed bets on every game according to the model's predictions or bet on games where the model meets a specified confidence threshold. This threshold will be another tunable model parameter.

I am interested in repeating the process above using player stats instead of team stats to build and compare model performance. The absence of a team's best player generally affects an NBA team more than teams in other sports, which would be reflected in the betting lines. Models built using team stats would not pick up on this, but building a model using player stats from the players in each game's lineup would better account for this.

References

[1] Nate Silver. *The Signal and the Noise: Why so Many Predictions Fail — but Some Don't*. Penguin Books, 2012.

[2] David Purdum and Ryan Rodenberg. *Future of Sports Betting: The Marketplace*. May 2017. URL: http://www.espn.com/chalk/story/_/id/17892685/the-future-sports-betting-how-sports-betting-legalized-united-states-the-marketplace-look-like.

[3] *NBA Regular Season League Standings - 15-16*. URL: http://www.covers.com/pageLoader/pageLoader.aspx?page=/data/nba/standings/2015-2016/sortable/standings_wins.html.