Beating the Odds
Betting on the NBA with Data

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 perception that anyone can do it 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 successful sports bettor. Oddsmakers are very good at their jobs and teams' records against the spread 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. For example, if a given team is favored by 5 points and wins by more than 5, they have "covered" and won 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 a number 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

I chose basketball as a testbed for two main reasons. The first reason is that basketball is arguably the most deterministic of the major American professional sports, providing an easier first problem to tackle than other sports do. 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

The second reason I chose to look at the NBA instead of other sports is that a wealth of NBA stats are readily available on their website stats.nba.com. The site exposes a wide variety of information in JSON format through various 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 so far this 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 for this task.

4 Outline

The first step is to build a classification model to predict NBA game winners given the strengths of the two teams in question. The model will be trained with games from previous seasons and tested against games already played this season. The second step involves building models to predict winners against the spread and over/unders. I am interested in repeating the first two steps using player stats instead of team stats to build the models and comparing the performance of the models. 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. I will also experiment with how much data to use to predict a given game. Using an entire season's worth of stats to predict games is not a realistic scenario. Bettors will only have stats available from games up to a given date to aid decisions. 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 goes on and more data becomes available.

References

[1] Nate Silver. *The Signal and the Noise: Why so Many Predictions Fail - but Some Don't*. Penguin Group, 2012. ISBN: 978-0-14-312508-2.

[2] David Purdum and Ryan Rodenberg. *Future of Sports Betting: The Marketplace*. 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.