Using Gaussian Processes to Predict NBA Game Outcomes

Andrew Cukierwar

May 2018

1. Introduction

In the NBA, a team's success depends on the number of games that they win in the regular season and the playoffs. Predicting the outcomes of each game that a team plays can therefore be a useful tool for the team. Furthermore, predicting the outcomes of games has applications in the world of sports betting. An excellent model that can accurately predict the margins of victory of games can make someone millions of dollars. The goal of this paper is to use Gaussian Process Regression to predict the outcomes of NBA games based on advanced NBA statistics

2. Methodology

Basketball can be broken down simply into offense and defense. A team's performance on the two sides of the court are best measured by offensive rating and defensive rating, which represent points scored and points allowed per 100 possessions. Dean Oliver, a basketball statistician, determined through regression models that a team's offensive and defensive ratings can each be broken down into "four factors of basketball success" [1]. These four factors are Effective Field Goal Percentage, Turnover Percentage, Offensive Rebounding Percentage, and Free Throws per Field Goal Attempt. Together, the four factors explain about 99% of the variance in offensive and defensive ratings.

Effective Field Goal Percentage is similar to field goal percentage but weights 3-point field goals to be worth 1.5 times as much as 2-point field goals. Turnover Percentage measures how often a team turns over the ball per possession. This factor is important as it represents how often a team gets an opportunity to attempt a field goal or get to the free throw line. Offensive Rebounding Percentage represents how many offensive rebounds a team gets of available offensive rebounds. Offensive Rebounds are important as they essentially create extra possessions for a team, allowing an extra opportunity to score. The last factor, Free Throws per Field Goal Attempt, measures how often a team gets to the free throw line and how efficiently that team can make the free throws. This factor is important as free throws are more efficient than regular field goals. A player shooting 60% from the field is typically considered elite but a player shooting 60% from the free throw line is generally considered a poor free throw shooter. A table with the offensive and defensive four factors for the 2016-17 season is shown below. Note that all data was scraped from basketball-reference.com [2].

	Team	Pace	oEFG	oTOV	ORB	oFTR	dEFG	dTOV	DRB	dFTR
0	Atlanta Hawks	98.3	0.512	14.1	21.0	0.185	0.536	13.6	76.2	0.183
1	Boston Celtics	96.0	0.518	13.0	21.5	0.188	0.495	13.0	78.4	0.191
2	Brooklyn Nets	98.9	0.514	13.6	21.0	0.201	0.517	11.0	77.0	0.201
3	Charlotte Hornets	98.4	0.508	11.4	22.2	0.233	0.532	12.4	80.7	0.165
4	Chicago Bulls	98.3	0.497	12.6	20.6	0.164	0.542	12.4	80.6	0.184
5	Cleveland Cavaliers	98.0	0.547	12.6	20.1	0.214	0.540	12.2	77.3	0.166
6	Dallas Mavericks	95.6	0.513	11.6	18.0	0.166	0.532	12.9	78.9	0.193
7	Denver Nuggets	96.8	0.536	13.4	25.7	0.198	0.539	12.6	77.5	0.173
8	Detroit Pistons	96.2	0.512	12.3	22.7	0.169	0.524	13.7	78.5	0.172
9	Golden State Warriors	99.6	0.569	14.1	21.0	0.195	0.504	12.6	76.3	0.186

Since the goal of this paper to predict the outcomes of games using advanced statistics, the advanced statistics shown above must be converted into features to be passed into a model. For each game played, there is a home team and an away team, each with four offensive and four defensive factors, meaning there are 16 factors in total. For each factor, I used the following formula to convert those 16 factors into 4 features:

(Home Team Offensive Factor + Away Team Defensive Factor)
- (Away Team Offensive Factor + Home Team Defensive Factor)
= Feature

This formula works because the home team would hope that the first two statistics are maximized and the last two are minimized. Therefore, subtracting the last two from the first two creates a feature that is properly correlated with the offensive and defensive aspects of each factor.

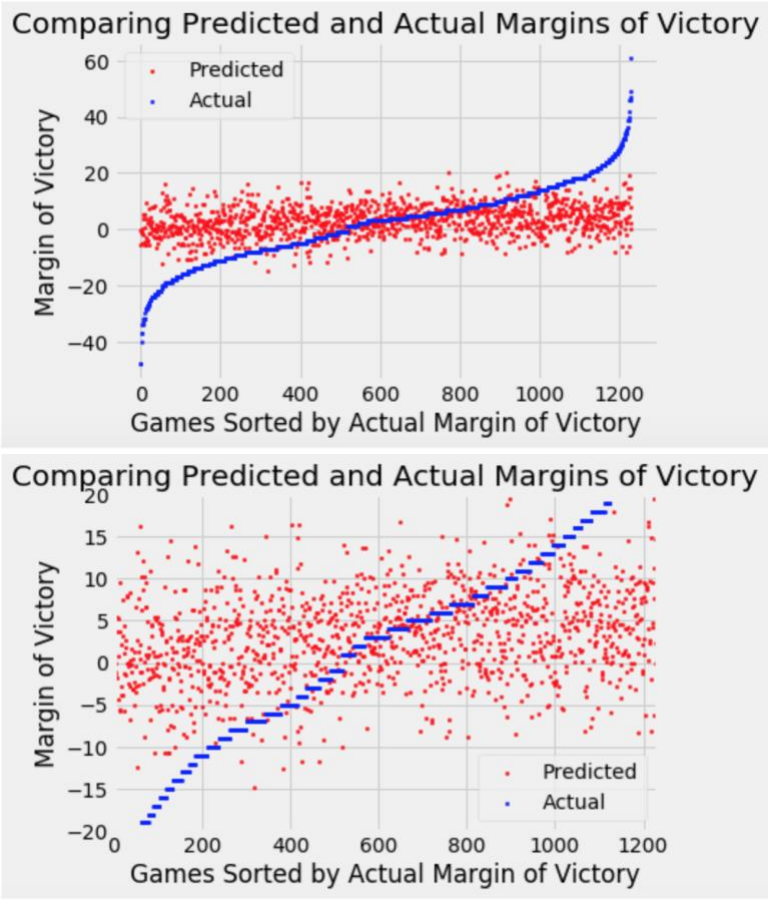
Once the features are set for each game, we can begin passing the data into the model to predict the margin of victory in terms of the home team. So, the margin of victory is positive if the home team wins and negative if the away team wins. A table showing game results for the first games of the 2016-17 season is shown below.

	Date	TeamAway	PointsAway	TeamHome	PointsHome	MarginHome
0	Tue Oct 25 2016	New York Knicks	88	Cleveland Cavaliers	117	29
1	Tue Oct 25 2016	San Antonio Spurs	129	Golden State Warriors	100	-29
2	Tue Oct 25 2016	Utah Jazz	104	Portland Trail Blazers	113	9
3	Wed Oct 26 2016	Brooklyn Nets	117	Boston Celtics	122	5
4	Wed Oct 26 2016	Dallas Mavericks	121	Indiana Pacers	130	9
5	Wed Oct 26 2016	Houston Rockets	114	Los Angeles Lakers	120	6
6	Wed Oct 26 2016	Minnesota Timberwolves	98	Memphis Grizzlies	102	4
7	Wed Oct 26 2016	Charlotte Hornets	107	Milwaukee Bucks	96	-11
8	Wed Oct 26 2016	Denver Nuggets	107	New Orleans Pelicans	102	-5
9	Wed Oct 26 2016	Miami Heat	108	Orlando Magic	96	-12

The model I used to predict the margins was a Gaussian Process Regression model with a mean of zero and a squared exponential covariance function with ell = e^7 and sigma = e^2. I performed a manual grid search on several covariance functions with varying hyperparameters to determine the covariance function and hyperparameters that I used. I then trained the model on the features and margins of victory of every game from the 2016-17 season.

3. Results

After the model was trained, I tested the model on games from the 2017-18 season by predicting the margin of victory for every game in the season. In the scatter plots below, I plotted the predicted margins of victory and the actual margins of victory, sorting by actual margin of victory. The second scatter plot is zoomed in from the first scatter plot to better show the correlation between the red and blue dots.



As we can see in the scatter plots, the predicted margins themselves are not very accurate. However, the predicted outcomes of the games, in terms of wins and losses, is fairly accurate. In the two plots above, more so apparent in the second plot, when the actual margin of victory is above zero, most of the predicted margins of victory are also above zero. This means that the model generally predicted a win when the actual outcome was a win.

In terms of predicting wins and losses, the model predicted 782 of 1230 games correctly, which comes out to around 63.6%. While this percentage may not seem that high, it is actually a respectable percentage when compared to various other methods of prediction. A simple coin flip would predict the right outcome 50% of the time. Picking the home team every time would predict the right outcome 58% of the time due to home court advantage. Experts who set betting lines in Las Vegas predict the right outcome 69.8% of the time [3]. Seeing as those experts likely use far more data and also use in-season data, my model predicting only 6% less wins appears to be acceptable.

4. Playoff Simulation

I decided to attempt to apply my model by simulating the playoffs last season and this season using my model's predictions. For each matchup, I predicted the margin of victory for the home team and the margin of victory when the teams play at the away team's court. These two predicted margins of victory are not equal as the model implicitly accounted for home court advantage. I then used a table from the Sports Book Review's website [4] to convert that predicted margin to a win probability. I then used a random number generator to determine which team won each game.

For the 2017 playoffs, my simulation had the Celtics playing against the Raptors in the Eastern Conference Finals instead of playing against the Cavaliers. In the Western Conference Finals, my simulation had the Golden State Warriors playing the Houston Rockets instead of the San Antonio Spurs. While my model correctly had the Warriors advancing to the finals, my model had the Raptors advance to the finals in the east even though in actuality they never made it past the second round. My model also correctly had the Warriors winning it all in a short series. So, the only real difference is my model had the Raptors replacing the Cavaliers in the Eastern Conference Finals and the Finals with both of those series projected to end in around the actual number of games it took.

To predict the 2018 playoffs, which are still currently going on, I retrained the model using data from the 2017-18 season rather than the data from the 2016-17 season. I ran my simulations again, which correctly predicted all of the winning teams of the first round matchups. In the second round, which is currently being played in real life, my model projects that the Raptors beat the Cavaliers in 7 games, the Celtics beat the 76ers in 7 games, the Jazz beat the Rockets in 7 games, and the Pelicans sweep the Warriors. The first two outcomes are definitely realistic outcomes and the third one is possible but unlikely. The last outcome is impossible at this point as the Warriors won the first two games of the series already. My simulation ended with the Raptors beating the Pelicans in the finals, which is skeptical as both teams are unlikely to make it past the current round.

5. Discussion and Conclusion

Three main factors hindered the performance of the model. First and foremost, the original approach to try to predict games in the 2017-18 season purely based off outcomes and statistics from the 2016-17 season is not ideal. While predicting 63.6% of games correctly at the beginning of the season is pretty significant, using in-season data could have made the predictions even more accurate. Likely some sort of weighting between 2016-17 season data and in-season data would have been ideal. This also accounts for the fact that rosters and coaches change between seasons. Second, the model performed well for only using the four factors, but to truly increase the performance, more features could have been used. Features that take into account injuries, travel, playing back to backs, and more. Lastly, training the model on data from the playoffs in previous years would have helped to improve playoff predictions. Some teams, such as the current Cavaliers, play better in the playoffs because they prioritize their whole season around the playoffs. They do not put as much stock in regular season games as some other teams might. All in all, I am satisfied with the accuracy of the model. It predicted almost 64% of games correctly purely based on statistics from the previously season and it only used a few features. For quite a simple model, it seemed to do well at predicting wins and losses.

References

[1] https://www.basketball-reference.com/about/factors.html
[2] https://www.basketball-reference.com
[3] https://hackernoon.com/how-to-predict-the-nba-with-a-machine-learning-system-written-in-python-part-ii-f276b19520b9
[4] http://www.bettingtalk.com/win-probability-percentage-point-spread-nfl-nba/