Multivariate linear regression: A data driven approach to predicting midseason NBA point spreads

Mackenzie Qu

10 December 2020

Abstract

|From the squeak of shoes on the court to the klaton of a shot clock buzzer, the excitment of basketball lies within its fast and unperdictable nature, which gradifies a massive crowd of gamblers. For the purpose of predicting pointspread outcome for any midseason NBA games, multiple regression models are developed using team's previous game stats from the same season. The model shows an 83% accuracy within 3 points, which is then tested using randomly selected midseason games from 2015-2018. Our model for pointspread is also applicable with regards to predicting the Money Line outcome, which may provide an overview of the upcoming season betting lines. | |Keywords: basketball, Bayesian logistic regression, multivariate linear regression, NBA, pointspread

Introduction

Basketball has gained millions of passionate fans over the years, and its intensity often extends beyond the courts into various bets. As the NBA remains to be the best basketball league in the world, the betting market of it also holds significant value. The NBA is composed of 30 teams; each team plays 83 games for a regular season which provides 1230 opportunities for gamblers to cash in. The most popular basketball bet is the Money Line, in which bettors simply choose a desired team to win. A more intriguing bet on point spread, however, has been selected to be the main focus in this report as it also includes the Money Line results. Point spread is when one bets on a team to win by a specific number or not lose more than that number. In this paper, we experimented with both multivariate linear regression model and Bayesian regression model with R Core Team (2020), hoping to predict a profitable betting line on the spreads of NBA games.

For the purpose of this paper, actual NBA game statistics(Kelepouris (2018)) from 2014 to 2018 are used to assist the model development. The data includes all games in a season, along with game stats such as field goal percentage and 3 point percentage generated by each team. We have chosen to use the the average of field goal percentage, 3 point percentage, free throw percentage, total rebounds, assists, and turnovers of both team as predictor variables. In addition, we have also used a Bayesian logistic regression to predict the general outcome for each game. Our model for point spread has a residual standard error of 3.8, which is inadequate in accuracy in comparison to the betting lines. However, the Bayesian model for predicting the winning team appears to be significantly accurate through verification.

The result could potentially be used to predict point-spread in the upcoming NBA season. However, the model is not sufficient enough to achieve the accuracy of the NBA betting lines chosen by professional oddsmakers. Our model could be significantly improved by adding the actually in-game statistics as a predictor. The rest of the report proceeds as follows. Section 2 describes the historical NBA data we have used to develop and verify the models, including some adjustments made. In section 3, we demonstrate various models including multivariate linear regression and Bayesian regression, aiming to provide statistical reasoning behind the results. Section 4 shows the result of the models and verifies the model performance through randomly selected games and their betting lines respectively. In which the results will be discussed thoroughly in section 5, as well as some potential weaknesses and future work. # Data

Model

field goal percentage~team+opponent+avg field goal percentage+Opp avg

point spread(team points-opponent points) ~ team + opponent_team + predicted_field_goal_percentage(past game avg home or opp) + predicted_opponent_field_goal_percentage + predicted_3point_percent + predicted_opponent3point_percent

Result

Discussion

Apendix

Reference

Kelepouris, Ionas. 2018. *NBA Team Game Stats from 2014 to 2018*. https://www.kaggle.com/ionaskel/nba-games-stats-from-2014-to-2018/metadata.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.