NBA 2018-19 Season Playoff Prediction

Basketball playoff prediction team

December 15, 2021

1 Introduction

Sports are interwoven with human civilization as an essential entertainment to society. From wrestling, the first sport we can prove in human history in 15,300 years ago (Pledge Sports, 2020), to over 800 types of sports like basketball, football, soccer in today's society (Topend Sports, 2021). Sports have become an essential element to most people's lives; some sports further built up a professional and structural league for the players and viewers for various reasons, like economic benefits or exchange of technical skills. In these leagues, a season often starts with predicting the teams that will enter the playoff and the championship team. The value behind successfully predicting an uncertain outcome is enormous, which can be a game result, a team can enter playoff season, or a team will play the championship game. From the perspective of personal betting, people can win times of their odds for successfully predicting an uncertain outcome, winning an amount of money that can be a person's total earnings for years. From the perspective of the club management team, an accurate prediction of their team can enter the playoff or not can help the management to make an annual plan. The management team can implement strategies to save or earn a significant amount of money. Clubs gain revenue from game tickets and products like player jerseys, but the amount varies by the team performance and the game record. For example, the management team can sign a lot of expensive and skillful players to build a competitive team that can win the championship game, so that the club can gain much revenue from tickets and products even they paid a lot for players' salaries. Alternatively, they can sign high-value players that are good but not great to save salary costs, which the clubs would not lose money even the players are not good enough to make playoff season.

In National Basketball Association (NBA), Bob Voulgaris was the director of quantitative research and development of Dallas Mavericks. He worked as a professional in sports betting, especially in NBA games, used the NBA play-by-play data to make a computer program for NBA betting, which the program worked perfectly and helped him gain a mysterious amount of money. In 2018, the owner of Dallas Mavericks, Mark Cuban, who appreciates his capability of accurately predicting the result of NBA games, hired him as the director of quantitative research and development of Dallas Mavericks (Steve, 2021).

The story of Bob Voulgaris provides evidence of how important is an accurate prediction to those professionals and clubs.

1.1 Research Question

Inspired by the story of Bob Voulgaris, we want to build a model that can predict the probability of each teams winning the NBA final championship. In this project, we will use regular season performance to predict the winner in the NBA playoffs. There are two main components to make the prediction. First, we will use Bayesian logistic regression to fit a model that can predict the probability of team i beats team j. Second, we will use these probabilities to simulate the outcome of the NBA playoffs using binomial random variables. It allows us to make prediction about the final champion.

The two metrics of the regular season performance were used. They are offensive efficiency (OEFF) and defensive efficiency (DEFF). We processed the data to get the difference of each value between every two teams. The prior distribution is based on the difference of offensive efficiency and the difference of defensive efficiency. Then, we further processed the prior distribution with the Markov chain Monte Carlo (MCMC) sampling method to obtain a full model and two reduced models. As a result, our full model predicted the 2019 champions Toronto Raptors have about 7.6 percent chance to win the NBA championship.

2 Dataset

In this project, two publicly available datasets are downloaded from websites and processed to form our final dataset. First, the regular season performance dataset contains the (2012-13 - 2018-19) regular season team performance metrics. The dataset is extracted from the website "NBA Stuffer" (nbastuffer.com). The data contain detailed rankings and performance metrics of teams for every regular season. Available data on these teams can be dated back to 2007 (Ughr, 2007). To better determine the likelihood update's effect in posterior results, predictors offensive efficiency (OEFF) and defensive efficiency (DEFF) of each team in regular seasons are selected for likelihood update. According to definitions listed in "NBA Stuffer", OEFF represents number of points scored per 100 possessions on each team, and DEFF represents number of points allowed per 100 possessions (Ughr, 2007).

The (2013 - 19) NBA playoff results are extracted from the website (basketball-reference.com). The extracted data contains the results of every playoff games played in the corresponding years. To train a model that predicts the winner team in playoffs, we decides to use the regular season data from 2012-13 to 2017-18 as the training set and focus on predicting the playoffs in 2018-19 data.

2.1 Processing Data

To identify the difference of team's ability in i-th playoff games, difference of OEFF (ΔO_i) and difference of DEFF (ΔD_i) between the two teams are calculated using their OEFF and DEFF from the regular season performance data of corresponding year. For example, in i-th playoff games, team t's OEFF and DEFF from the corresponding season will be subtracted with team t''s OEFF and DEFF from the corresponding season to get ΔO_i and ΔD_i using the regular season performances over the 2012-2013 to 2018-2019 season. Calculations on ΔO_i and ΔD_i are based on the playoff match agendas among the teams from 2012-2013 season to 2017-2018 season using each team's OEFF and DEFF in the regular season dataset.

In addition to identify ΔO_i and ΔD_i, results of playoff matches are referenced based on the playoff agendas and actual outcomes. Here, $y_i = 1$ represents team t beats team t' in the i-th playoff game while $y_i = 0$ represents team t' beat team t. Considering the representation of team t and team t' are relative towards each other, it is expected that inverse numbers on the outcome of ΔO_i and ΔD_i are presented in the data and in the following models. All these mentioned variables are used to form a basic logistic regression model for training and likelihood updates. Further details on the usage of the model is shown in the section of Model.

2.2 Notice on Data Usage

The project chooses to utilize the data starting from the 2012-2013 season. Also, due to the 2020 global COVID-19 pandemic, the regular season and playoff season of 2019-2020 and the ongoing 2020-2021 were significantly disrupted and they may result in a different trend of team performances from previous seasons. This makes the 2018-19 season the last full season without external influence and thus this team decides to utilize relevant data from 2012 to 2019.

As for the prediction of posterior result, matching sequences (i) and 16 involving teams (t or t', where both represents two competing teams in i-th playoff game) demonstrated in the 2019 playoff data will be followed for generating the posterior result. Based on NBA playoff rules, these 16 playoff teams consist 8 teams from the east conference and 8 teams from the west conference (2019 NBA Playoffs Bracket, 2019). The final match involves two teams from each conference to compete. Thus, simulations on the outcome of the final match will be used for posterior predictions through using regular season data and playoff data.

3 Model

The project believes that the binary characteristic of the response variable (0 for losing and 1 for winning) is most suitable for a logistic regression model. The model is roughly expressed as

$$\log(\frac{\mathbb{P}(y_i = 1)}{\mathbb{P}(y_i = 0)}) = X_i \times \beta$$

where X_i is the predictor matrix with variables ΔO_i and ΔD_i. Here X_i has dimension 502×2, for 502 rows of train data and 2 columns of predictors. The matrix β represents the regression coefficient and records the log-odds of winning probability when the predictors – delta offensive rate and delta defensive rate – change by 1 unit. The dimension of beta is 2×1, where vector $\beta[1, 1]$ is the regression coefficient for delta offensive rate and vector $\beta[1, 2]$ for delta defensive rate.

Theoretically, the intercept term α represents a baseline for success probability for it considers the situation when OEFF and DEFF are both 0. The project chooses to discard the intercept term for simplifying home team and away team variation. Without the intercept term, we can swap the OEFF and DEFF values interchangeably when modeling the situations of away team versus home team.

3.1 Prior Distribution

The regression coefficients β is the prior that needs to be specified before training the model. As the coefficient is expressed as a 2×1 matrix in the logistic regression model here the project uses β_1 to represent $\beta[1, 1]$ and β_2 to represent $\beta[1, 2]$. To determine the 2 priors, the project decides to examine the 2 corresponding variables independently. For ΔO_i, the project expects the difference of offensive rate between 2 teams to make relatively huge difference in winning probability; for ΔD_i, the project expects that the increase in defensive rate differences may not correspond to a positive change in winning probability.

After tuning the prior for several times, the project chooses a positive mean for β_1 and negative mean for β_2 because we expect that changing ΔO_i for 1 unit will correspond 0.7 times change in log-odds; similar for β_2 that we expects a change of roughly -0.4. For the standard deviation, the project expects the variance of priors to be small for the sake of accuracy. As a result, the project fits the prior $\beta_1 \sim N(0.75, 0.05)$ and $\beta_2 \sim N(-0.39, 0.05)$.

Since both 2 predictors have approximate range [-9, 9], the project decides not to do standardization for variables and coefficients.

To test the prior predictive model's performance, the project computes 2 prior predictive plots. The plots fit the prior predictive probabilities on 2 grids ranging from -9 to 9 for both ΔO_i and ΔD_i. The log-odds is calculated by $X_i \times \beta \forall x \in$ x_grid and the prior predictive probability on the y-axis is transformed using the inverse logit function

$$\text{inv\_logit}(w) = \frac{1}{1 + e^w}$$

.

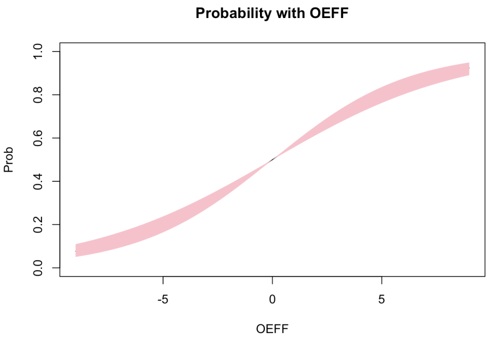

Figure 1: Prior prediction of ΔO_i

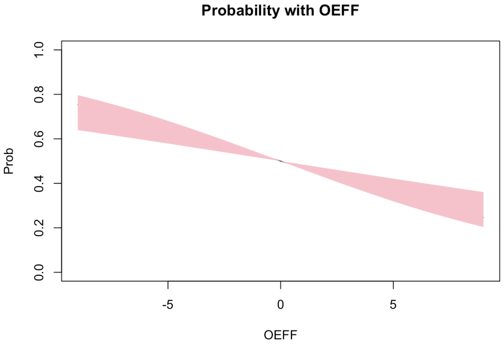

Figure 2: Prior prediction of ΔD_i

For β_1 and ΔO_i, we can see that the probability shows a positive association on [0, 1] while ΔO_i changes from -9 to 9 on the x-axis. The prior prediction for β_2 is negatively correlated with the winning probability and shows a relatively flatter trend. The range of probability is approximately between 0.2 to 0.8 in the second plot.

3.2 Modeling with Stan

Markov chain Monte Carlo (MCMC) is a sampling method that can estimate a probability distribution without knowing all model properties. The MCMC takes a random walk through the given data points for all possible values and constructs chains to approximate the sample from the posterior. As random walk has the property of not deciding the next step only based on the current position, data processed by the Stan program is expected to eventually converge. Since the model consists of 2 predictors, we decide to transform each variable into matrix form. The project uses Stan to implement the Metropolis-Hastings Algorithm.

The Stan codes consist of 4 main sections. The first section "data" is the input of the model, where the project determines the dimensions of the matrices, the predicting and response variables, and the grids for posterior predictions. The Stan codes have model structure $y_i = \begin{bmatrix} \Delta O_i & \Delta D_i \end{bmatrix} \times \beta$, or:

$$\begin{bmatrix} y_1 \\ y_2 \\ .. \\ y_{502} \end{bmatrix} = \begin{bmatrix} X_{1,1} & X_{1,2} \\ X_{2,1} & X_{2,2} \\ .. & .. \\ X_{502,1} & X_{502,2} \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

where y_i is a 502×1 matrix and X_i is a 502×2 matrix with 2 columns representing the predictor.

The second section "parameter" specifies the regression coefficient β and the third section "model" inputs the prior values for β and creates the Bernoulli logistic model for y_i. The order or multiplication is fixed since that we are doing matrix multiplication.

The last section "generated quantities" generates probability grids for posterior prediction results. Since matrix X_i consists of 2 predictors, the ideal plot is 3D. To visualize the outcome in 2D plots, the project decides to fix one predictor on certain values and put the other predictor on the x-axis. To show the posterior predictive plots properly, the project decides to fix one predictor at its mean, mean minus one standard deviation, and mean plus one standard deviation. The rough idea of posterior predictive probabilities look like

$$\text{log-odds}(y_i) = X_1 \times \beta_1 + \mu_{X_2} \times \beta_2$$

$$\text{log-odds}(y_i) = X_1 \times \beta_1 + (\mu_{X_2} - \sigma_{X_2}) \times \beta_2$$

$$\text{log-odds}(y_i) = X_1 \times \beta_1 + (\mu_{X_2} + \sigma_{X_2}) \times \beta_2$$

where μ_{X_2} is the mean of predictor 2, in this case ΔD_i, and σ_{X_2} represents the standard deviation for predictor 2.

3.3 Playoff Simulator

The playoff simulator is an R function that can simulate the outcome of the entire playoff bracket. The function requires 2 inputs. First, a list of the 16 playoff team names. The order of the team names is important. The first 8 teams are grouped into one conference as well as the remaining 8 teams. In each conference, the teams are further grouped into 4 pairs of 2 by order, which should correspond to the first-round schedule of the playoff. The second required input is a named pair-wise probability matrix of one team that beats another. The names in the two inputs need to be matched. For simulating the “best-of-7” series between two teams in each pair, the function will find the probability of “team 1” beats“team 2” using the pairwise-probability matrix. The order of the teams within pairs of two does not matter. The corresponding probability p will be used in a $Binomial(7, p)$ random variable to simulate the game results. If the random variable is greater than or equal to 4, then “team 1” wins. Otherwise, if it is less than 4, then “team 2” wins. The teams will first compete within the conference until only one team is left for each conference. The remaining two teams will play against each other for the final championship. The function will return a list of 2 team names, where the first name is the simulated final champion and the second name is the team that lost in the NBA finals.

3.4 Two Reduced Models

We followed a similar procedure to fit the logistic regression model on offensive efficiency(OEFF) and defensive efficiency(DEFF) separately and observed that the 2 reduced models performed worse than the full model, and the reduced model with DEFF performed better than the reduced model with OEFF based on their prediction on the probability of Toronto Raptors win in 2019 (Full model: 7.6%, reduced model with OEFF: 4.5%, reduced model with DEFF: 6.1%). Although we expected that OEFF would be a stronger prediction in winning probability than DEFF in our prior setting, the prediction of Toronto’s winning probability is higher than that of OEFF. However, the overall predictions of a reduced model with only DEFF were quite unreasonable given that it gave very high winning probabilities to Utah Jazz and Oklahoma City Thunder when they didn’t win in the first round. In that aspect, both the full model and reduced model with only OEFF performed relatively well because their highest winning probabilities were given to the two teams (Golden State Warriors and Milwaukee Bucks) that achieved second and third place at the end.

4 Results

4.1 Posterior coefficients

To get the posterior value of β_1 and β_2, we used RStan to extract 4000 β_1 and β_2 using MCMC. The result shows that the mean value of β_1 for all these 4000 samples is 0.305 and the mean value of β_2 is -0.244. As we tune the prior of β_1 and β_2, the trend of posterior regression coefficients does not change a lot. This is coherent with the Bayesian characteristic that the posterior probability will be close to the likelihood if given sufficient data. In our case, we have 502 training data, which can be considered significantly large. The project has also found out that the posterior $|\beta_1|$ is almost always larger than that of $|\beta_2|$, meaning that predictor ΔO_i has a stronger effect on the winning probability.

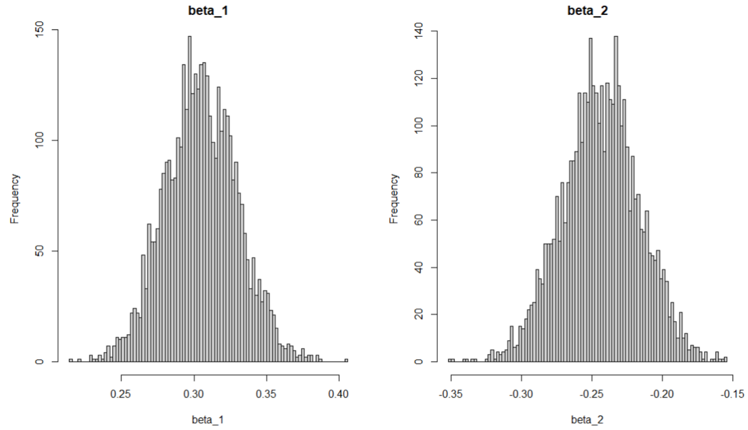

Figure 3: plot for beta coefficients

4.2 Posterior Model

We use all 4000 β_1's and β_2's as coefficients of the logistic regression. For the first plot, we fix deltaDEFF and compute the probability of deltaOEFF. Here we can see that the x-axis deltaOEFF is positively associated with the posterior probability, where the black line is predictor deltaOEFF when deltaDEFF is at its mean, the grey line is deltaDEFF at mean minus one standard deviation, and the red line is deltaDEFF at mean plus one standard deviation. We can see the grey line generally has a higher probability. This might be because that the two predictors tend to be oppositely plotted, which makes sense in real life as the offensive rate is calculated as the opposite of the other team's defensive rate.

For the second plot, we fix deltaDEFF and compute the probability of deltaOEFF.We use same colors as the first plot. The x-axis deltaOEFF is negatively associated with the

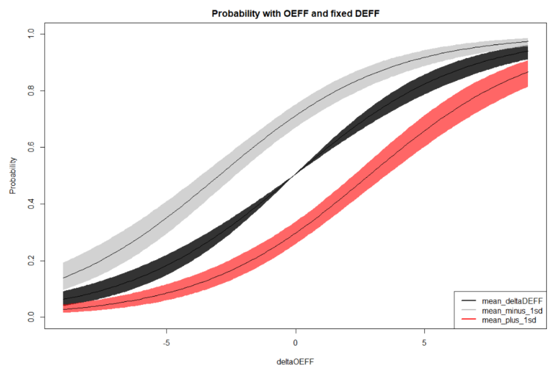

Figure 4: Posterior prediction of ΔO_i

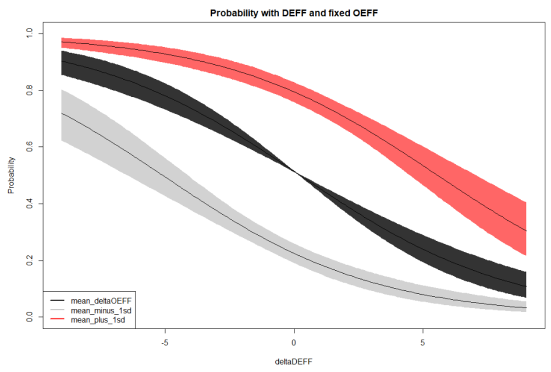

Figure 5: Posterior prediction of ΔD_i

posterior probability. The position of grey and red lines also makes sense as the defensive rate is calculated as the opposite of the other team's offensive rate.

4.3 Simulation

Teams	Probabilities
Boston	0.00725
Denver	0.00725
Golden State	0.23575
Houston	0.02875
Indiana	0.00025
Milwaukee	0.60175
Oklahoma City	0.00075
Philadelphia	0.00100
Portland	0.01825
Toronto	0.08957
Utah	0.00925

Table 1: Simulation result

The table above shows the predicted probabilities for each team to win the 18-19 season championship. The predictions are made by running 4,000 simulations using the 4,000 pair-wise probability matrix of one team beats another calculated using the 4,000 posterior samples of (β_1, β_2). The mean squared error $(y - \hat{p})^2$ on the 18-19 playoff results is 0.213, where $\hat{p}$ is the posterior mean predictive probability of team 1 beats team 2. y equals to 1 if team 1 beats team 2 and 0 otherwise. Flipping a fair coin will achieve MSE of 0.25.

	Full Model	Reduced Model 1	Reduced Model 2
Training	0.2440	0.2555	0.2486
Testing	0.2460	0.2568	0.2492

Table 2: MSE for Model Comparison

4.4 Cross Validation

To compare the three models, we conducted 5 fold cross-validation experiment for each models on the 12-18 dataset. We computed the MSE $(y - \hat{p})^2$ on both training and testing sets. Table 2 shows the result of the cross validation.

Here we can see that the mean square error for each model is very close to the baseline 0.25, which is a good performance. Among the 3 models, the full model performs better than the 2 reduced model. This may show that predicting with both delta offensive rate and delta defensive rate is a better choice than using only one predictor.

5 Future Discussions

5.1 Strength and Potential Weakness

The models failed to consider the replacement and injuries of teammates, the occurrence of accidents, and the shift between home team and away team. In real-life scenarios, however, a sports match can almost never be fully predicted by a mathematical model. What the project can do here is to increase model accuracy by acknowledging more factors.

6 Conclusion

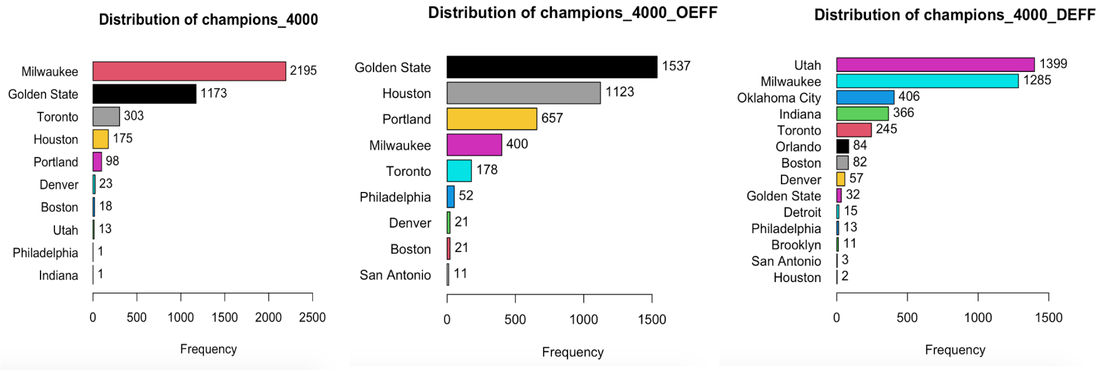

Figure 6: winning probability

Our full model predicted that the winning probability of Toronto Raptors would be 7.6%, which was higher than the two reduced model. Although winning probability of the true winner was not high, the full model did a great job in giving both Milwaukee Bucks and Golden State Warriors relatively high winning probability (60% and 23.5%), which corresponded to the result of the 2019 competition in which they achieved the second and the third place. The reason that our model generated very low winning probability for Toronto Raptors and high winning probability for Golden State Warriors is that we didn't consider the randomness of real-life scenarios. For instance, Durant tore his right Achilles tendon on June 10, 2019 in Game 5 of the NBA Finals against the Toronto Raptors. The MSE of our model is 0.21, which indicates that the forecast result is relatively close to the actual result. Therefore, it is reasonable to conclude that offensive rate and defensive rate are plausible factors to consider when predict the winners.

References

Ughr, Serhat. (2007). *About NBAstuffer.* NBAstuffer. https://www.nbastuffer.com/about-nbastuffer/

Ughr, Serhat. (2007). *Team Evaluation Metrics.* NBAstuffer. https://www.nbastuffer.com/analytics-101/team-evaluation-metrics/

2019 NBA Playoffs Bracket. (2019). Retrieved December 12 2021, from ESPN, ESPN website. http://www.espn.com/nba/bracket/_/year/2019

Oldest sports in the world. Pledge Sports. (2020, June 9). Retrieved December 13, 2021, from https://www.pledgesports.org/2018/02/oldest-sports-in-the-world/.

Complete list of sports from around the world. Topend Sports, science, training and nutrition. (n.d.). Retrieved December 13, 2021, from https://www.topendsports.com/sport/list/index.htm.

Steve. (2021, June 9). *Bob Voulgaris.* LegalSportsBetting.com. Retrieved December 13, 2021, from https://www.legalsportsbetting.com/famous-sports-bettors/bob-voulgaris/.