Predicting NBA Wins with Bayesian Machine Learning

Kevin Malloy, Nick Clifford, Brigitte Hogan

Problem Description

Each year, 30 NBA teams compete over 82 separate games. Knowing how well a team will perform can influence personnel decisions like trades, hirings, or firings, as well as the expected revenue of the team. Consequently, both team managers and fans are highly invested in team wins. What's more, the May 2018 Supreme Court ruling on the Professional and Amateur Sports Protection Act (PASPA) as unconstitutional opened the door for increased legality in sports gambling and the involvement of oddsmakers. Having the ability to predict team outcomes more accurately than oddsmakers would result in a competitive advantage that could be used to turn a profit. Even small improvements in the ability to predict a team's performance could have major financial implications.

The goal of our project is to use Bayesian statistics to model all teams' win percentages given their pre-season win projections and their in-season results through a specific time period. We chose our in-season results cutoff date to be December 5th, which corresponds to about 25% of the total season already having been played. We will model a team's performance both with point estimates of projected wins and credible intervals that determine uncertainty. Both elements are crucial, as they provide an estimate of results and a measure of how certain these results are.

Mathematical Linkage

Our data consists of pre-season win estimates for each team for the 2016-2020 seasons, the in-season records for each team up until December 5th for the 2016-2020 seasons, and the final records for each team for the 2016-2019 seasons. The final records for each team in the 2020 season are obviously not yet available because the season is still being played. This data is perfect for a beta-binomial conjugate prior analysis approach (Figure 1A). The prior data are the teams' pre-season projected win percentages. The binomial likelihood can be modeled with the total number of games played through December 5th, with the total wins through the same date being the number of "successes". The posterior will take the form of a beta distribution that incorporates both the prior and the likelihood. The Maximum A Posteriori (MAP) estimate shows what percentage of games a team is expected to win. We also employ the use of a hyperprior on the parameters of the beta distribution to express uncertainty about the values of alpha and beta, which correspond to the pre-season projections. We analyze how this hierarchical model compares to the results that do not utilize a hyperprior.

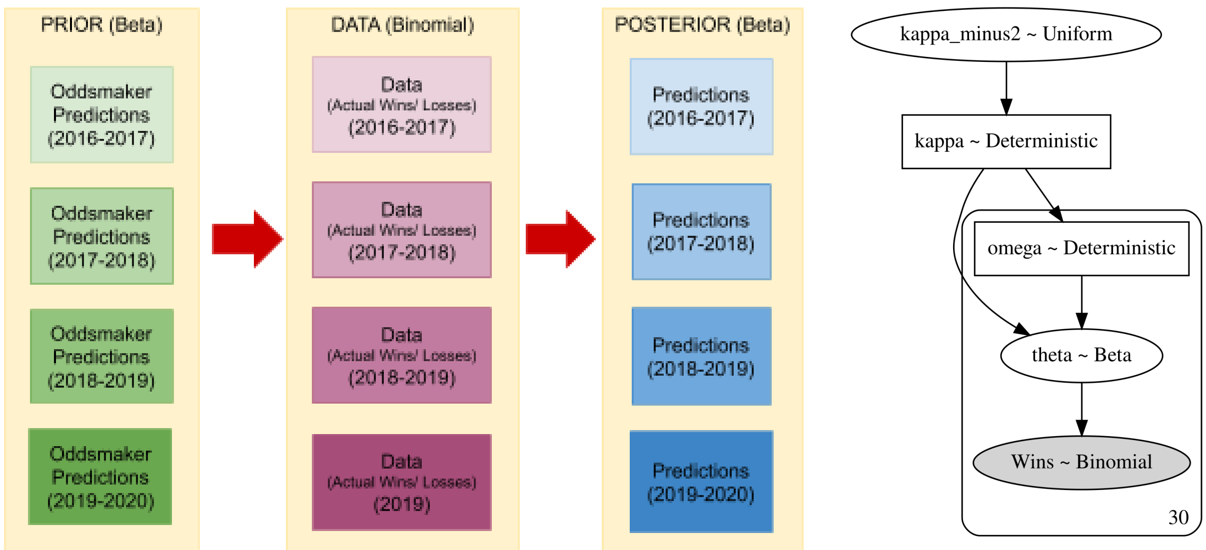
The relevant equations linking the beta binomial model to our data are shown below.

$F(\theta|y) = Beta(\alpha + y,\ \beta + n - y)$, the form of the beta posterior given the data

$E = \frac{\alpha + y}{\alpha + y + \beta + n - y}$, the expected value of the MAP estimate of the posterior where α and β are the beta prior parameters, and *n* and y are the number of games played (n) and won (y) by a team through December 5th, respectively.

Figure 1. Graphical Models

A. Conjugate Approach B. Hierarchical



Bayesian Methods

<u>Conjugate Prior Approach</u>

The most difficult part of our beta-binomial model is determining what values to give the alpha and beta parameters of the prior. In other words, we must decide how informative or non-informative to make the prior. We experiment with using a noninformative prior, a prior that is informative but not overpowering of the binomial portion of the data, and a very strongly informative prior. Having the previous seasons' results will allow us to compare how different prior weights perform on the data. For the informative priors, we plan to have the mode of the beta distribution be equal to the pre-season projected win percentage.

As a first approach, we set the prior to be noninformative with an alpha and beta of one (see NBA_A_ConjPrior.ipynb). This setting assumes that each team has an equal likelihood of finishing the season with any given win percentage. Obviously this assumption is not very reasonable, but we wanted to use this noninformative approach to compare to our models with more informative priors. This method resulted in the widest posterior distributions of the three non-sampling attempts, which is intuitive because the non-informative prior corresponds to more uncertainty.

Next, we fit a model with highly informative priors. We set the alpha value to the number of wins a team was projected out of 82, and the beta value to their projected number of losses. This method fits the mode of the distribution at their estimated win percentage. Because only about 20 games of the season have been played up until our cut off point, such high prior parameter values carry very substantial weight in the posterior distribution. This result can be observed by both the equation for the posterior and the expected value of the posterior in the previous section. Because of the strong weight of the prior, the posterior is more narrow, and less uncertain, than the other options.

Finally, we fit another estimate with informative priors, but the alpha and beta parameters were scaled down by a factor of 0.4. This maintains the ratio of alpha and beta, but makes the distribution less informative. This method hopes to find a happy medium between non-informative priors and very strong priors that could crowd out the data in the likelihood function.

Hierarchical Model Approach

As opposed to manually setting a scaling value to weight the prior knowledge, we also created a hierarchical model to specify a hyperprior on our prior parameters (see NBA_B_Hierarch.ipynb). Specifically, the certainty of our preseason projection should be inversely proportional to the amount of games that have been played to our specified cutoff point. In other words as the season gets closer to its end, it should make sense that the prior makes less of an impact on our posterior estimate.

$$\alpha = \omega(\kappa - 2) + 1$$

$$\beta = (1 - \omega)(\kappa - 2) + 1$$

In order to do create this model, we redefine the α & β parameters above in terms of the mode ω and "certainty" κ, where the certainty represents the sample size of our prior estimate and the mode can be solved for given our projections as the distribution's mean. We place a hyperprior on κ to represent it as a random variable from a uniform distribution (Figure 1B).

In order to solve for the posterior MAP estimates, we resort to using gradient-based MCMC NUTS sampling offered in PyMC3. We run sampling 4 chains drawing 100,000 samples each and take into account a burnin of 100 samples (see NBA_C_Overlay.ipynb). We then use a kernel density estimator that calculates the estimator bandwidth based off of Scott's Factor, which is used to compare the posterior distributions between this approach and the conjugate prior approach.

Results & Conclusions

To evaluate our different models, we compare both the maximum a posteriori estimates and the credible intervals for each team's prediction on the previous years' results. The greatest difference between the MAP and results occur with the noninformative prior (Table 1). The moderate prior and strong prior approach yielded very similar results for the accuracy of their MAP estimates.

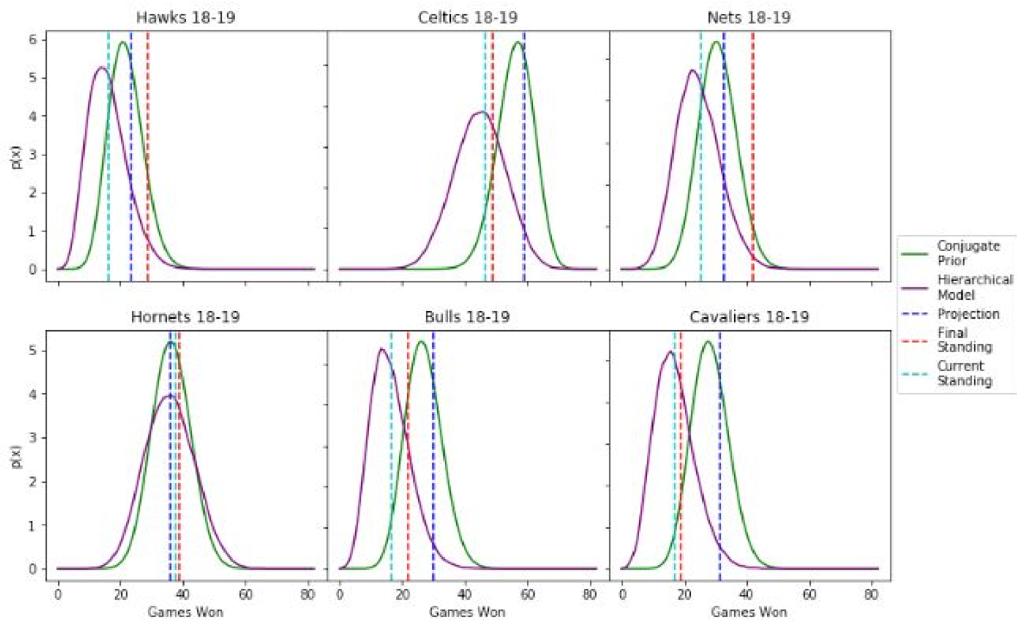
Table 1. Comparison of the different prior approaches.

Method	Median MAP - Result Difference	Percent of Test Data Results within 90% CI
Noninformative	6.17	95.5%
Moderate Prior	4.40	93.3%
Strong Prior	4.32	80.0%

Regarding credible intervals, the goal of our analysis is to compute a prior weight such that 90% of the previous years' actual results fall within the 90% credible intervals from our posterior distributions. For our strong prior model, only 80% of the results did, meaning that the credible intervals were not wide enough. For the noninformative approach, 95.5% of results did, meaning that the intervals were too wide. The moderate prior approach was closest with 93.3%, which was confirmation to us that it was the best model of the three.

The choice of prior can be framed in the context of a bias-variance tradeoff. Selecting a non-informative prior results greater variance (less confidence in the results), and could lead to a less biased estimate. On the other hand, a strong prior results in more confidence (less variance), but could result in a more biased estimate.

Figure 2. Hierarchical and moderately scaled conjugate approach comparison, 2018-2019 season



Overall, the hierarchical model follows the current standings of the data while the conjugate approach more closely follows the preseason projections (Figure 2). This is due to the hyperprior set on the certainty parameter of the number of games left in a season, thereby reducing the weight of the prior knowledge. The hyperprior results in a larger range of confidence at the expense of reducing the probability density around the MAP estimate. This model performed close to that of a noninformative prior, with a median difference between the MAP and Final result being 6.02 games. For future work, it may be worthwhile to improve the distribution of the certainty prior, test the results of an ensemble model that is the combination of both the conjugate and hierarchical models, and factor in substantial personnel changes to teams during the season.

Works Cited

"Standings by Date." *Basketball Reference*, www.basketball-reference.com/leagues/NBA_2019_standings_by_date_eastern_conference.html.

Abrams, Evan. "2019-20 NBA Win Total Tracker: How All 30 Teams Are Progressing: The Action Network." *Action Network*, The Action Network, 2 Dec. 2019, www.actionnetwork.com/nba/nba-win-totals-betting-odds-2019-2020.

Chavez, Chris. "NBA Team 2017-18 over/under Records." *Sports Illustrated*, 30 Aug. 2017, www.si.com/nba/2017/08/30/nba-over-under-win-totals-all-30-teams-2017-2018-season.

Foxsports. "2016-17 NBA Season Win Total Odds and Predictions." *FOX Sports*, 20 Sept. 2016, www.foxsports.com/nba/story/2016-2017-nba-season-win-total-odds-and-predictions-092016.