Predicting NBA All-Stars Using Bayesian Logistic Regression

Timo Pew

December 10, 2019

1 Introduction

Being selected to play in the All-Star game is one of the highest forms of individual recognition an NBA player can receive. Fair or not, discussions about a player's legacy is largely predicated on All-Star selections. More recently, All-Star selections even have financial implications, with many teams including conditional bonuses in contracts based on making the All-Star team. Though the intent of the game is to recognize and showcase the 24 best players in the league of the given season, every year there appears to be players deserving of the recognition that do not receive it. We believe that there may be additional variables contributing to the selection of All-Star rosters beyond a player's performance in a given season.

To test our hypothesis and further understand what factors are related to All-Star selection, we have data for each player in each season that played more than 300 minutes from the 2008-2009 through the 2018-2019 seasons. For each player and season, in addition to whether or not the player made the All-Star roster, we have an assortment of variables of two different types: variables that directly measure season-specific performance and variables that are unrelated or indirectly related to performance. Performance variables include season averages of 'box-score' statistics such as points (PTS), rebounds (REB), assists (AST), steals (STL), and blocks (BLK) per game as well as more advanced metrics that attempt to quantify the overall impact a player makes such as value over replacement player ($VORP$) and win shares per 48 minutes ($WS48$). The non-performance variables include whether or not they previously made an All-Star roster (PAS), the proportion of games their team has won (Win), and the number of TV households (TV) in the market they play in. If those selecting the All-Star rosters are truly choosing the 24 best players in a given season, these variables will have a minimal effect on the probability of being selected and not help predict future All-Star selections.

The primary purpose of our analysis is to determine if the non-performance variables have an effect on the probability of a player making the All-Star game. We use a fully Bayesian approach to logistic regression to carry out our analysis. A secondary purpose of our analysis is to compare the results of our Bayesian analysis with results obtained through a standard frequentist approach. In the following sections we specify our model for both the Bayesian and frequentist analyses, detail our method to draw from the posterior distribution of the unknown parameters, assess the convergence of our posterior draws, and present the results of our analysis.

2 Methods

We assume that the dichotomous All-Star selection outcome (where a 1 indicates being selected and 0 indicates not being selected) for a given player and season, denoted Y_i, are independent Bernoulli random variables with probability mass function $P(Y_i = y_i|\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$ for $y_i \in \{0, 1\}$, $0 < \pi_i < 1$ for $i = 1$, $\ldots$, n where $n = 4206$. We define the relationship between π_i, the probability of being selected, and the explanatory variables as follows:

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta} \quad \text{for } i = 1, \ldots, n. \tag{1}$$

where $\boldsymbol{x}_i'$ contains p explanatory variables for the i^{th} observation, β_0 represents the log-odds of making the All-Star team when $\boldsymbol{x}_i$ is the zero vector, and $\boldsymbol{\beta} = [\beta_1 \ldots \beta_p]'$ contains the effects of the p variables on the probability of making the All-Star team. We standardized the continuous variables of interest because the common scale makes setting the same priors on each coefficient reasonable, simplifying the prior elicitation process. We will primarily consider two sets of explanatory variables. Our baseline model includes the performance variables PTS, REB, AST, and $WS48$, as well as the non-performance variables PAS, Win, and TV. We will also consider a model with only the performance variables.

We do not have substantial prior knowledge or intuition about the parameters β_0, β_1, $\ldots$, β_p because of their complicated interpretations. We suspect that many of them have a positive effect on the probability of making the All-Star team, but we have little intuition as to the location or variation of the parameters so we set normal priors on them. Because we do not know the location and scale of the coefficients, we put hyperpriors on their mean and precision. In our fully Bayesian framework, we chose the following priors on β_0, β_1, $\ldots$, β_p with their associated hyperpriors:

$$
\begin{aligned}
\beta_j|\mu_j,\tau_j &\overset{ind}{\sim} \mathcal{N}(\mu_j,\tau_j) \quad j=0,1,\ldots,p \\
\mu_j &\overset{iid}{\sim} \mathcal{N}(0,b) \quad j=0,1,\ldots,p \\
\tau_j &\overset{iid}{\sim} Gamma(c,d) \quad j=0,1,\ldots,p
\end{aligned}
\tag{2}
$$

We set normal hyperpriors on the mean and gamma hyperpriors on the precision of $\beta_0, \beta_1, \ldots, \beta_p$ because they have the correct support and we can take advantage of conditional conjugacy to sample from the posterior distribution more efficiently as we will demonstrate. For our primary analysis we set diffuse hyperparameters $b = 0.1$, $c = 1$, and $d = 1$ where b is the precision of μ_j. We set the mean of each μ_j to be zero as something of a null hypothesis that the variables have no effect on making the All-Star team. We also used values of $b = 64$, $c = 40$, and $d = 20$ to test how sensitive our analysis was to our chosen priors. These choices do not necessarily reflect an alternate specification that we consider to be good choices for the hyperparameters, but purely as an exercise to understand how sensitive our analysis is to our chosen parameterization.

Under the assumed sampling model and prior specification, the joint posterior density of the parameters $\beta_0, \ldots, \beta_p$; $\mu_0, \ldots, \mu_p$; and $\tau_0, \ldots \tau_p$ is proportional to:

$$
\begin{aligned}
&\prod_{i=1}^{n}\left(\frac{exp(\beta_0+\boldsymbol{x}_i'\boldsymbol{\beta})}{1+exp(\beta_0+\boldsymbol{x}_i'\boldsymbol{\beta})}\right)^{y_i}\left(\frac{1}{1+exp(\beta_0+\boldsymbol{x}_i'\boldsymbol{\beta})}\right)^{1-y_i} \times \prod_{j=0}^{p}\sqrt{\tau_j}\,exp\left(\frac{-\tau_j}{2}(\beta_j-\mu_j)^2\right) \times \\
&\prod_{j=0}^{p} exp\left(\frac{-b}{2}\mu_j^2\right) \times \prod_{j=0}^{p}\tau_j^{c-1}exp(-\tau_j d), \text{ where } n = 4206
\end{aligned}
\tag{3}
$$

Because the joint posterior density has an unknown form, we must use a simulation based approach to approximate the posterior density. Our particular selection of priors and hyperpriors allows us to take advantage of conditional conjugacy to sample more easily from the posterior distribution. The full conditional density of of each parameter can be found using Equation 3. The full conditional densities are proportional to the joint posterior of all parameters. The full conditional density of an individual β_j, μ_j, and τ_j for $j = 0, 1, \ldots, p$ are shown below.

$$
\begin{aligned}
&p(\beta_j|\cdot) \propto \left(\prod_{i=1}^{n}\frac{exp(x_{ij}\beta_j y_i)}{1+exp(\beta_0+\boldsymbol{x}_i'\boldsymbol{\beta})}\right) exp\left(\frac{-\tau_j}{2}(\beta_j-\mu_j)^2\right) \\
&p(\mu_j|\cdot) \propto exp\left(\frac{\tau_j+b}{2}\left(\mu_j-\frac{\beta_j}{1+b/\tau_j}\right)^2\right) \implies \mu_j|\cdot \sim N\left(\frac{\beta_j}{1+b/\tau_j}, \tau_j+b\right) \\
&p(\tau_j|\cdot) \propto \tau_j^{c-1/2} exp\left(-\tau_j\left(\frac{(\beta_j-\mu_j)^2}{2}+d\right)\right) \implies \tau_j|\cdot \sim Gamma\left(c+\frac{1}{2}, \frac{(\beta_j-\mu_j)^2}{2}+d\right)
\end{aligned}
\tag{4}
$$

As shown in Equation 4, the full conditional density of μ_j and τ_j for $j = 0, 1, \ldots, p$ are proportional to the kernel of Normal and Gamma random variables respectively. We take advantage of this property and implemented a Metropolis-within-Gibbs sampling scheme to get samples from the posterior distribution. The algorithm to obtain a single chain of draws from the posterior distribution is given in Algorithm 1.

It is important to note that our choice of a Normal proposal distribution for $\beta_0, \beta_1, \ldots, \beta_p$ is symmetric, thus making the updates for $\beta_0, \beta_1, \ldots, \beta_p$ Metropolis updates rather than Metropolis-Hastings updates. We chose the proposal standard deviation, σ, to be 0.5 after trial and error. As is often the case, we worked on the *log* scale for the Metropolis updates for computational efficiency. Thus M_r becomes $l(\beta_j^{(i)*}|\cdot) - l(\beta_j^{(i-1)}|\cdot)$ where $l(\beta_j|\cdot) = ln(p(\beta_j|\cdot))$, and $\beta_j^{(i)}$ is set to $\beta_j^{(i)*}$ with probability $e^{(\min(0,M_r))}$. To get starting points in step 1 of the algorithm, we first drew random values for μ_j and τ_j from their prior distribution, after which we used these values to sample from the prior distribution of β_j. We chose to run 5 chains of 10,000 iterations with 1,000 burn-in iterations for a total of 45,000 iterations. We assessed the convergence of our 5 chains based on the trace plots and $\hat{R}$. We also calculated the acceptance rate and effective sample size to further understand the mixing of our simulated draws and give confidence to the validity of our posterior inference.

Algorithm 1: Metropolis-within-Gibbs

1. Choose starting points for β_j, μ_j, and τ_j ($\beta_j^{(1)}$, $\mu_j^{(1)}$, $\tau_j^{(1)}$) for $j = 0, 1, \ldots, p$
2. For i $= 2, \ldots, B$:
   (a) For j $= 0, 1, \ldots, p$:
      (i) Draw a new proposed value, $\beta_j^{(i)*}$ from $\mathcal{N}(\beta_j^{(i-1)}, \sigma)$, where σ is a tuning parameter
      (ii) Compute the Metropolis ratio, $M_r = p(\beta_j^{(i)*}|\cdot)\,/\,p(\beta_j^{(i-1)}|\cdot)$
      (iii) With probability $\min(1, M_r)$, set $\beta_j^{(i)} = \beta_j^{(i)*}$, otherwise set $\beta_j^{(i)} = \beta_j^{(i-1)}$
   (b) Draw $\mu_j^{(i)}$ from its full conditional distribution, $\mu_j|\cdot \sim N\left(\frac{\beta_j^{(i)}}{1+b/\tau_j^{(i-1)}}, \tau_j^{(i-1)} + b\right)$ for $j = 0, 1, \ldots, p$
   (c) Draw $\tau_j^{(i)}$ from its full conditional distribution, $\tau_j|\cdot \sim Gamma\left(c + \frac{1}{2}, \frac{(\beta_j^{(i)} - \mu_j^{(i)})^2}{2} + d\right)$ for $j = 0, 1, \ldots, p$

To assess the fit of the model, we used a posterior predictive check. We simulated All-Star selections using our posterior draws from β_0, β_1, …, β_p and the observed explanatory variables. For each simulation (45,000 total) we calculated the number of All-Stars selected from each season. The posterior predictive p-values are the proportion of simulated number of All-Stars in each season that surpassed the actual number of All-Stars selected that season. A good model fit would result in a similar number of simulated All-Stars as the true number of players selected for each year, with a p-value near 0.5. A p-value near 1 indicates a season where the simulated data resulted in far more All-Stars than actually selected, whereas p-values near 0 indicate the opposite. This test also helps us see if there is potentially a year-to-year trend that we are not capturing. For example, there is a general consensus that the NBA has been getting increasingly talented and the games are being played at a faster pace. This could lead to more players appearing 'worthy' of selection over time. This trend would be reflected in low p-values for earlier years and high p-values for later years. We also calculated the posterior predictive p-values using the alternative prior specification ($b = 64$, $c = 40$, and $d = 20$) to determine how sensitive our analysis was to our prior specification.

We used two different methods to determine to what extent non-performance variables effect selection to the All-Star game. First, we considered properties of the marginal posterior distributions of the coefficients associated with non-performance variables. Positive coefficients indicates that there is a greater probability of being selected to the All-Star team when the associated variable increases. We constructed 95% confidence intervals and point estimates for the regression coefficients using the profile likelihood in a frequentist analysis, and compared them to the posterior mean and 95% credible intervals from the Bayesian analyses under both prior specifications. To further understand the effect that these variables have on All-Star selection we calculated the posterior predictive probability of being selected (π) for different values of each non-performance variables holding everything else constant. We did this holding two different sets of performance variables constant. First, we set PTS to two standard deviations above the mean (20.9 points per game) and the other three to one standard deviation above the mean (6.5 rebounds per game, 4 assists per game, and 0.148 win shares per 48). Based on these values alone, we consider this the profile of a borderline All-Star player. Second, we set each of the performance variables to two standard deviations above the mean (20.9 points, 8.9 rebounds, and 5.8 assists per game, with 0.205 win shares per 48). We consider this profile very likely to be an All-Star. For both profiles we constructed the posterior predictive distribution of π_i under different settings of the non-performance variable of interest. In a frequentist setting there are no posterior predictive distributions. The best comparison we felt comfortable with was to compare the point estimates for the probability of being selected in the different settings.

Second, we compared the predictive abilities of the model including both performance and non-performance variables to the model with only performance variables. To do so, we held out the 2018-2019 season data and built both models on the remaining data. We used the marginal posterior distributions of the regression coefficients to predict the probability of each player in the 2018-2019 season being selected to the All-Star game. We predicted that a player made the All-Star roster when over 50% of their posterior predictive distribution of π_i exceeded 0.50. We also predicted whether or not each player in the 2018-2019 season made the All-Star roster based on a model built using data from the 2008-2009 to 2017-2018 season in a frequentist setting. We compared the confusion matrix from this analysis to the confusion matrices created using the posterior draws from our Bayesian model.

3 Results

Before performing posterior inference, we first assess the convergence of our 5 chains and examine other Markov chain Monte Carlo diagnostics. The trace plots of several parameters are shown in Figure 1. We overlapped the plots for each chain to visually compare them. In addition, we computed $\hat{R}$, the effective sample size, and acceptance rate for each parameter. These values for each β_j are displayed in Table 1. The effective sample size, $\hat{R}$, and acceptance rates for the μ_j and τ_j easily surpassed acceptable values so we do not display them here. It was unsurprising that the hyperpriors mixed easily since each proposed value is accepted in Gibbs sampling, whereas the Metropolis updating for the regression coefficients themselves do not always accept the proposed values. The only trace plots (shown in Figure 1) and effective sample sizes that raise concerns are those for β_0 and β_1. Both β_0 and β_1 struggled to mix well, with effective sample sizes of 451 and 505 respectively, although the acceptance rates were reasonable for both (0.255 and 0.168). Despite these moderate concerns, the $\hat{R}$ values fall well below the suggested threshold of 1.10, which means we do not have reason to believe the chains did not converge based on $\hat{R}$. It is possible that using a different standard deviation for the proposal distribution would have improved the mixing, however, we tried several values ranging from 0.3 to 0.8, of which 0.5 produced the best effective sample sizes and acceptance rates. Because β_0 and β_1 are also highly negatively correlated, perhaps using a multivariate update would have improved the mixing. While we would have preferred the trace plots to move around better with larger effective sample sizes, we feel comfortable with the mixing of our chains and confident that we have enough draws to make valid inference.

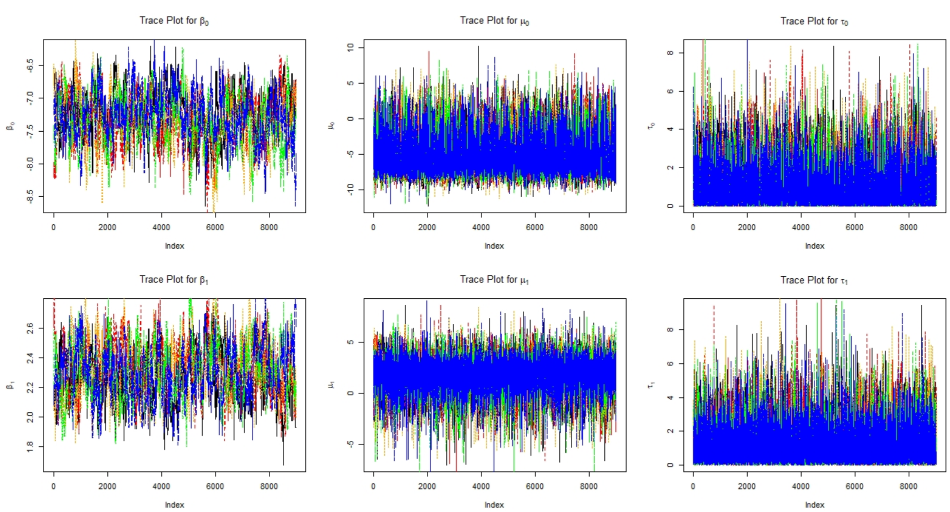

Figure 1: Trace plots for β_0, β_2, μ_0, μ_2, τ_0, and τ_2

Table 1: Effective samples, $\hat{R}$, and acceptance rate for each β_j

Diagnostic	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Effective Samples	451	505	1142	1548	1230	2795	1175	6314
$\hat{R}$	1.019	1.020	1.005	1.004	1.002	1.000	1.007	1.001
Acceptance Rate	0.255	0.168	0.162	0.167	0.215	0.333	0.256	0.247

To assess the fit of our model, we used the posterior predictive check explained in the Methods section. The posterior predictive p-values can be found in Table 2. The results from both parameterizations are fairly similar. The simulated number of All-Stars under the alternate specification was greater than the simulated number of All-Stars under our chosen specification overall, though in general, the simulated number of All-Stars selected each year was similar to the truth. This indicates that the data we observed is a reasonable realization under our posterior distribution and that there is not a noticeable year-to-year trend that our model does not capture. It is unsurprising that the prior parameterization seems to change the results somewhat since our alternative prior parameterization were stronger and remained centered around zero, while the data seem to indicate they are not close to zero.

Table 2: Posterior predictive test for both prior parameterizations

Parameterization	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
$b = 0.1$, $c = 1$, $d = 1$	0.572	0.082	0.596	0.275	0.150	0.509	0.054	0.475	0.912	0.402	0.906
$b = 65$, $c = 40$, $d = 20$	0.692	0.190	0.713	0.431	0.275	0.658	0.129	0.597	0.945	0.526	0.936

To determine the effects of the non-performance variables on All-Star selection, we first look at the characteristics of the marginal posterior distributions of the regression coefficients. Every single draw from the marginal posterior distribution of β_5 and β_6 were greater than zero, and 99.9% of the draws from the marginal posterior distribution of β_7 were greater than zero. While this does not necessarily prove that whether or not the player had previously been selected to an All-Star game, the proportion of games their team won, or the number of TV households in their market are causing the increased probability of making the All-Star game, we have strong evidence to support that there is a relationship between these non-performance variables and being selected to the All-Star game. Table 3 shows the upper and lower bounds of the confidence intervals and a point estimate for each regression coefficient for the Bayesian and frequentist approach. The point estimates and confidence bounds from our primary prior selection and a frequentist analysis are very similar. We feel this is the case because the sampling model is fairly simple and we set diffuse priors hyperpriors on the coefficients. We note that although the confidence bounds are similar, they have substantially different interpretations. The alternative specification resulted in the marginal posterior distributions of the coefficients being shrunk towards zero. This is unsurprising because under the alternative specification the distribution of the coefficients remained centered around zero while having the prior variability reduced.

Table 3: Point Estimates and Confidence Intervals for Regression Coefficients Using Bayesian and Frequentist Methods

Method	Type	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Bayesian (1)	Point Estimate	-7.301	2.306	0.679	0.684	0.393	1.535	1.025	0.320
	Lower 95% CI	-8.068	1.970	0.467	0.485	0.049	1.109	0.715	0.106
	Upper 95% CI	-6.610	2.658	0.896	0.874	0.728	1.970	1.332	0.526
Bayesian (2)	Point Estimate	-6.236	1.934	0.570	0.556	0.417	1.341	0.807	0.223
	Lower 95% CI	-6.848	1.621	0.364	0.363	0.117	0.946	0.537	0.027
	Upper 95% CI	-5.628	2.241	0.779	0.749	0.726	1.739	1.083	0.415
Frequentist	Point Estimate	-7.267	2.297	0.677	0.680	0.385	1.536	1.024	0.321
	Lower 95% CI	-8.075	1.954	0.459	0.485	0.049	1.108	0.720	0.110
	Upper 95% CI	-6.556	2.668	0.903	0.880	0.729	1.973	1.339	0.531

As explained, we also plotted the posterior predictive probability of being selected under several situations. The density estimates for the posterior predictive probability for each setting are in Figure 2. The two plots furthest on the right show the effect that previously being selected to an All-Star game has on being selected. For a player with the borderline All-Star profile, having previously been selected to an All-Star game improves the probability of selection from unlikely to likely. For a player with a strong performance profile, having previously made an All-Star roster improves the probability of selection from likely to almost certain. Team success has a similarly strong effect on the probability of being selected. As the win percentage increases, the probability of being selected increases substantially for both the borderline and strong All-Star profile. Market size has a less pronounced effect on the probability of being selected, but clearly still increases the probability of being selected. The dotted lines on each plot represents the frequentist point estimates of the probability of being selected in each setting. All of the point estimates are very close to the mode of the posterior predictive density, again highlighting how similar the results are for both approaches. At the same time, it also highlights how a Bayesian approach often expresses uncertainty more easily and intuitively. In summary, these plots provide visual evidence of the relationship between the non-performance variables and All-Star selection.

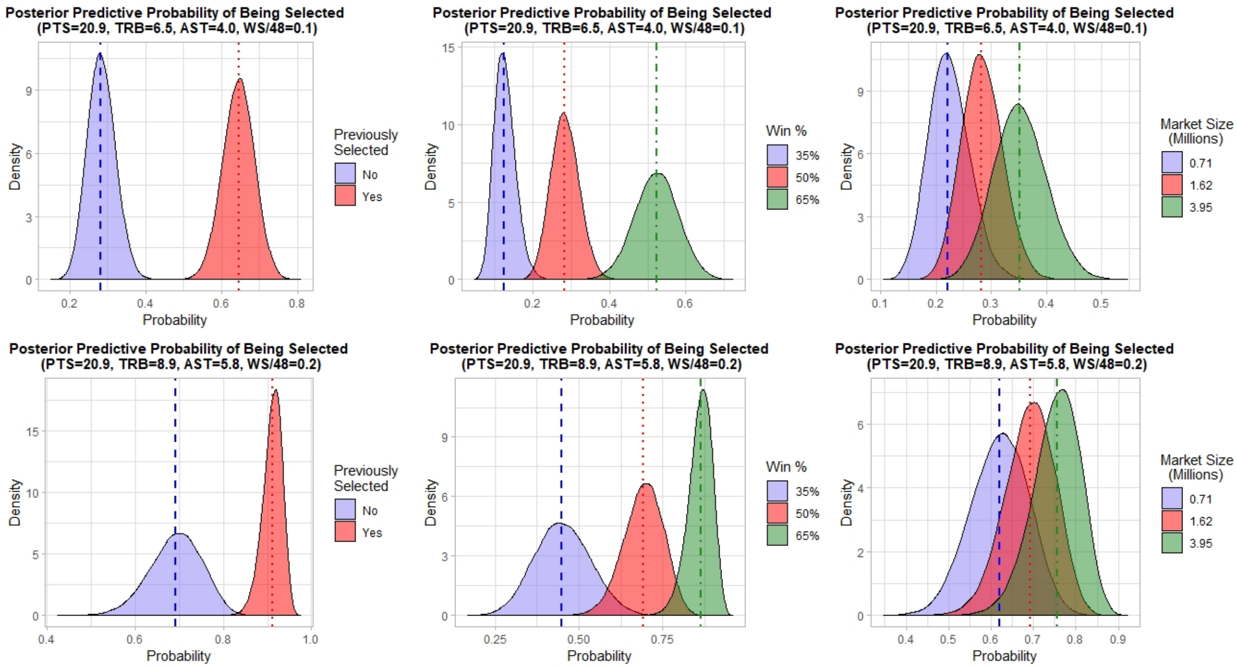

Figure 2: Posterior predictive probability of being selected under various conditions. The dotted lines represents the frequentist point estimates

As an additional assessment of the effect of non-performance variables on All-Star selection, we built a model including both types of variables and one with only performance variables holding out the 2018-2019 season. We used the posterior draws from both models to predict whether or not each player from the 2018-2019 season made the All-Star game. The confusion matrices for both models are in Table 4 with the model using only performance variables on the left and the model using both types on the right. The model that includes non-performance variables correctly predicts an additional 3 All-Stars, while also only incorrectly predicting 5 players as being All-Stars that were not selected. We feel these are very good predictions, especially when considering some context behind the 2018-2019 All-Stars. Two players, Dwyane Wade and Dirk Nowitzki, were specially selected as additional All-Stars despite not contributing nearly enough to even be considered if they were not Hall-of-Fame worthy players in their final seasons (This also contributes to the theory that factors beyond performance in the season at hand affect selection). Removing these two leaves 23/25 accurately predicted. Using a frequentist approach to predict future All-Stars yielded the same results as Table 4. These predictions provide evidence that the non-performance variables have some practical significance in addition to statistical significance. That is, including them in our model increases the ability to predict of future All-Stars.

Table 4: Confusion matrix for model built with only performance variables on the left and for model built with both types of variables on the right using Bayesian approach

		Predicted	
		Positive	Negative
Truth	Positive	20	7
	Negative	10	366

		Predicted	
		Positive	Negative
Truth	Positive	23	4
	Negative	5	375

4 Conclusion

Using a Bayesian hierarchical logistic regression model, we found strong evidence that having previously been selected to an All-Star roster, team success, and the number of TV households in the local market have an effect on the probability of being selected to play in the NBA All-Star game. Not only were the marginal posterior distributions almost exclusively greater than zero, but a model including performance and non-performance variables predicted All-Stars for the 2018-2019 season better than one with only performance variables. In addition, we found that for this particular application, Bayesian and frequentist methods resulted in very similar results.

There were several shortcomings of the analysis. First, as we mentioned, our effective sample size was a little lower than we would have liked for β_0 and β_1. However, getting these two parameters to mix well was difficult even when using an off-the-shelf program (JAGS) to simulate draws from the posterior distribution. We feel comfortable that our implementation performed relatively well and we still felt confident we had enough samples to perform valid inference. Another shortcoming came from the data format. Each of the variables were collected at the end of the season, when in reality, the All-Star rosters are chosen about half way through the season. This makes it possible for some players to potentially look over or under qualified for the recognition due to large changes midseason. For example, a player who struggles in the first half of the season and plays at a high level afterwards is unlikely to make the All-Star team despite having the profile of an All-Star caliber player. Taking a snapshot of these variables at the time the rosters are announced may improve the model. From simply scanning through the players that were incorrectly predicted in-sample, it appears this may make some difference, although it does not appear to be substantial. We also have some concern about the interpretation of the effect that the number of TV households in the local market has on All-Star selection. In the NBA, big market teams often have more resources to acquire All-Star caliber players. Therefore, we exhibit caution assuming the effect of market size on All-Star selection is being driven by the player being in a big, popular market. Our analysis makes no attempt to define causality, therefore we still feel comfortable acknowledging that there is a relationship between market size and All-Star selection, whether the effect is being driven by being in a popular market or not. Similar to those concerns, we acknowledge that our choice of performance variables was a somewhat arbitrary choice. It is possible that the non-performance variables were in reality explaining some of the variation that would have been explained by other performance variables, had they been included.

In future research we would attempt to minimize the shortcomings just explained. We would like to explore a more robust method to more confidently attribute the observed effects of non-performance variables to the actual non-performance variables. Perhaps one strategy might be to initially include all possible variables and use a variable selection method to keep only the most important variables and see which non-performance variables remain in the model. In addition, we would like to consider models with more performance and non-performance variables than those we included. For example, a common theory is that the Eastern Conference is weaker and therefore easier to make the All-Star team. It would be interesting to explore if there is statistical validity to this hypothesis. It would also be of interest to determine which players have been most negatively impacted by a potentially biased selection process.