Prediciting Playoff Appearances of NBA Teams

APSTA 2123: Final Project

Bilal Waheed

5/16/2020

1 Introduction

The National Basketball Association (NBA) is a men's professional basketball league in North America. There are 30 teams within the league, each playing 82 games in a season. 16 teams can qualify to be in the playoffs, which extend the number of games played for those teams. Like many sports, the NBA maintains records of numerous metrics on a team and player basis. These metrics help understand the overall performance of each team.

In this analysis, we will be predicting which NBA team is likely to enter the NBA Playoffs based on aggregrate metrics. Examples include total points scored per game, total steals per game, free throw percentage, and offensive rebound percentage.

2 NBA Data Overview

The data were scraped from the ESPN website and consist of 30 teams over 10 seasons (2010-2019). The data contain 300 observations with several metrics for each team, and an indicator for whether or not that team made the playoffs in that season. Independence across teams and years is assumed.

For this analysis, the data are split into training and test sets, with the training set consisting of seasons 2010-2018, and the test set being the 2019 season. We will fit a model on the training data, and predict on the test data to evaluate performance.

Below is a sample of the dataset to understand and visualize its characteristics:

teams	PTS	FGM	FGA	FG_percent	X3PM	X3PA	X3P_percent
Milwaukee Bucks	118.1	43.4	91.1	47.6	13.5	38.2	35.3
Golden State Warriors	117.7	44.0	89.8	49.1	13.3	34.4	38.5
New Orleans Pelicans	115.4	43.7	92.2	47.3	10.3	29.9	34.4
Philadelphia 76ers	115.2	41.5	88.2	47.1	10.8	30.2	35.9
LA Clippers	115.1	41.3	87.5	47.1	10.0	25.8	38.8

FTM	FTA	FT.	OR	DR	REB	AST	STL	BLK	TO	PF	MOV	Playoffs
17.9	23.2	77.3	9.3	40.4	49.7	26.0	7.5	5.9	13.3	19.6	8.8	1
16.3	20.4	80.1	9.7	36.5	46.2	29.4	7.6	6.4	13.8	21.4	6.4	1
17.8	23.4	76.1	11.1	36.2	47.3	27.0	7.4	5.4	14.5	21.1	6.0	0
21.2	27.5	77.1	10.9	36.9	47.8	26.9	7.4	5.3	14.5	21.3	5.2	1
22.6	28.5	79.2	9.7	35.8	45.5	24.0	6.8	4.7	13.8	23.3	4.7	1

2.1 Generative Model

We will use a bernoulli data generating process to model the probability of a team being a playoff team or not. The model will incorporate the following variables as predictors:

- `FG_percent: Total number of shots scored divided by total shots attempted.`
- `REB: Total rebounds per game.`
- `AST: Total assists per game.`

These variables are basic metrics that assess overall team strength: `FG_percent` measures a team's scoring efficiency, `REB` measures how well can a team rebound the ball and not give the opposing team second-chance opportunities, and `AST` measures how well the team is passing and utilizing each member of the team.

The data are modeled in the following form:

$$logit(P_{Playoffs}) = \beta_0 + \beta_1 FGpercent + \beta_2 REB + \beta_3 AST$$

3 Priors and Drawing from Prior Predictive Distribution

3.1 Setting Priors

First we will determine the priors for each parameter of the model. We will use the `get_prior` function to see the coefficient names. This function will also help to give a starting point for setting prior distributions on the parameters. Results are shown below:

prior	class	coef	group	resp	dpar	nlpar	bound
	b						
	b	AST					
	b	FG_percent					
	b	REB					
student_t(3, 0, 10)	Intercept						

Based on the output, we will use Gaussian priors on the coefficients of the predictors. This is reasonable because, due to the dynamic structure and nature of NBA teams, a change in any metric could result in an increased or decreased likelihood of playoff appearance.

- **Field-goal percentage (FGP)**: Higher FGP generally corresponds to better shot selection, and more points, but that is not always the case. Therefore, this prior will be set to $N \sim (0.20, 0.40)$ to allow for the possibility that FGP could reduce a team's chance to be a playoff team.
- **Assists (AST)**: It is reasonable to believe that the higher assists could relate to overall team chemistry and ball-movement, which would increase a team's likelihood of being a playoff team. The prior on this will be set to $N \sim (0.20, 0.20)$.
- **Total rebounds (REB)**: This is a trickier metric and does not necessarily correspond to more points scored. This prior will be set wider to $N \sim (0, 0.50)$

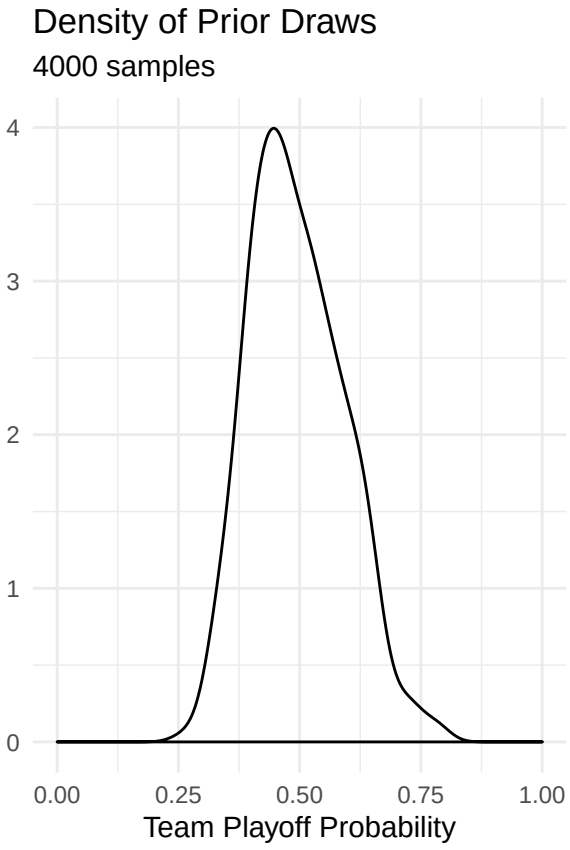
3.2 Prior Distribution Draws

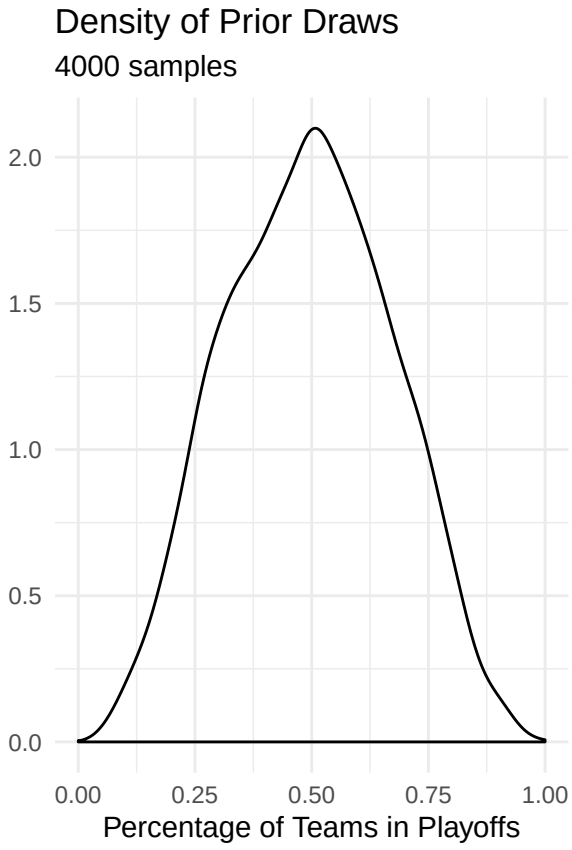
After setting these priors, we will now draw 4000 samples from the prior predictive distribution and examine the density plots of the prior draws to see if our priors are reasonable.

The code to set the priors and draw from the prior predictive distirbution is provided below:

```
# Set priors:
priors <- prior(normal(0, 1), class = "Intercept") +
  prior(normal(.20, 0.40), coef = "FG_percent") +
  prior(normal(0.20, 0.20), coef = "AST") +
  prior(normal(0, 0.50), coef = "REB")

# Create prior distributions:
prior <- brm(Playoffs ~ FG_percent + REB + AST, data = nbatrain, family = bernoulli,
             prior = priors, sample_prior = "only")

# Draw from prior distributions (4000 samples):
ppd <- pp_expect(prior)
```

The density plots of the predicted outcome below provide two different perspectives on the prior predictive distribution: the left shows the average of 4,000 draws of the predicted probability of being in the playoffs for each team, and the right shows the total estimated percentage of teams being in the playoffs for each draw.





The prior predictive distribution shows that teams roughly have a 50% chance to make the playoffs, on average, with no team having exactly 0% or 100% chance. Additionally, on average, 50% of total teams will make the playoffs. These priors make sense, because every season there are 16 out of the 30 teams that make the playoffs.

4 Posterior Distribution: Conditioning on Observed Data

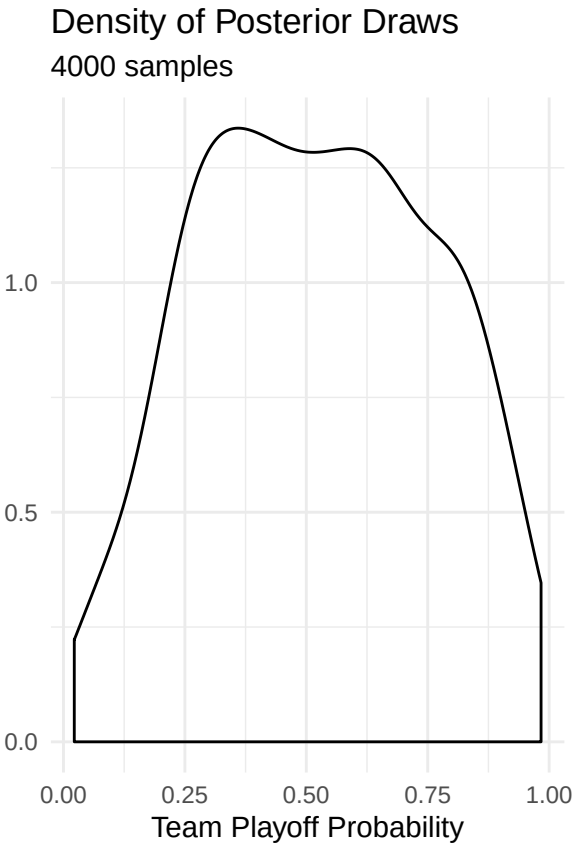
Now that are priors are in good shape, we will examine the posterior distribution after conditioning on the observed data. The code and output are provided below.

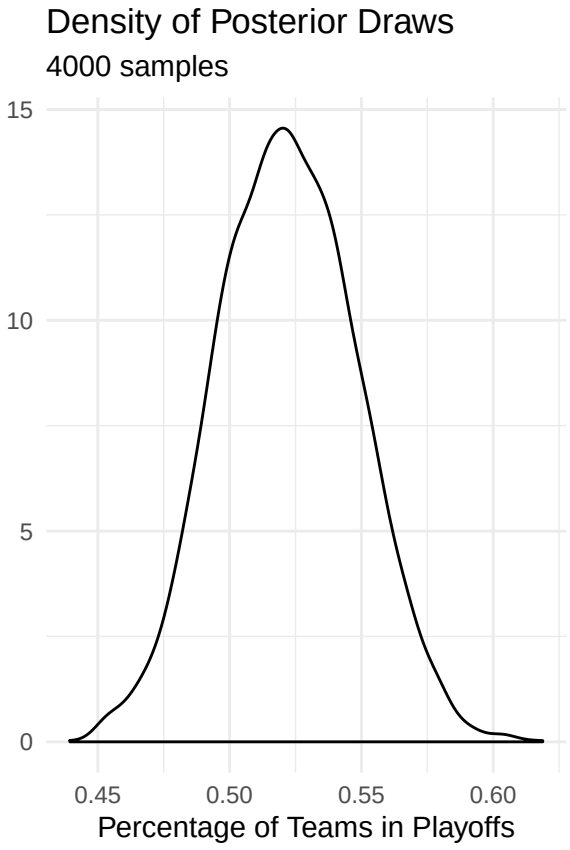
```
# Posterior distribution:
post <- update(prior, sample_prior = "no")

# Draw from posterior distribution (4000 samples):
post_draws <- pp_expect(post)
```

Shown below, after conditioning on the data, we can see adequate model convergence with Rhat values equal to 1.00 as well as large effective sample sizes. The updated parameter estimates are consistent with our prior beliefs about the effect of those variables. It is interesting to note that the estimate for `AST` is close to 0, indicating that higher assist values have little to no impact on the likelihood of being in the playoffs.

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	-45.74	6.71	-59.20	-33.04	1	3925	3051
FG_percent	0.74	0.12	0.52	0.97	1	3714	3288
REB	0.26	0.08	0.10	0.41	1	3951	3331
AST	0.06	0.08	-0.10	0.22	1	3799	3037

The below density plots of the posterior distribution also show reasonable estimates of playoff probability. The plot of percentage of playoff teams has narrowed to include the range from approximately 0.45 to 0.60, with a mean right around 0.52. This is consistent with the actual data, because each season, there are exactly 16 out of the 30 teams (53.33%) that qualify for the playoffs.



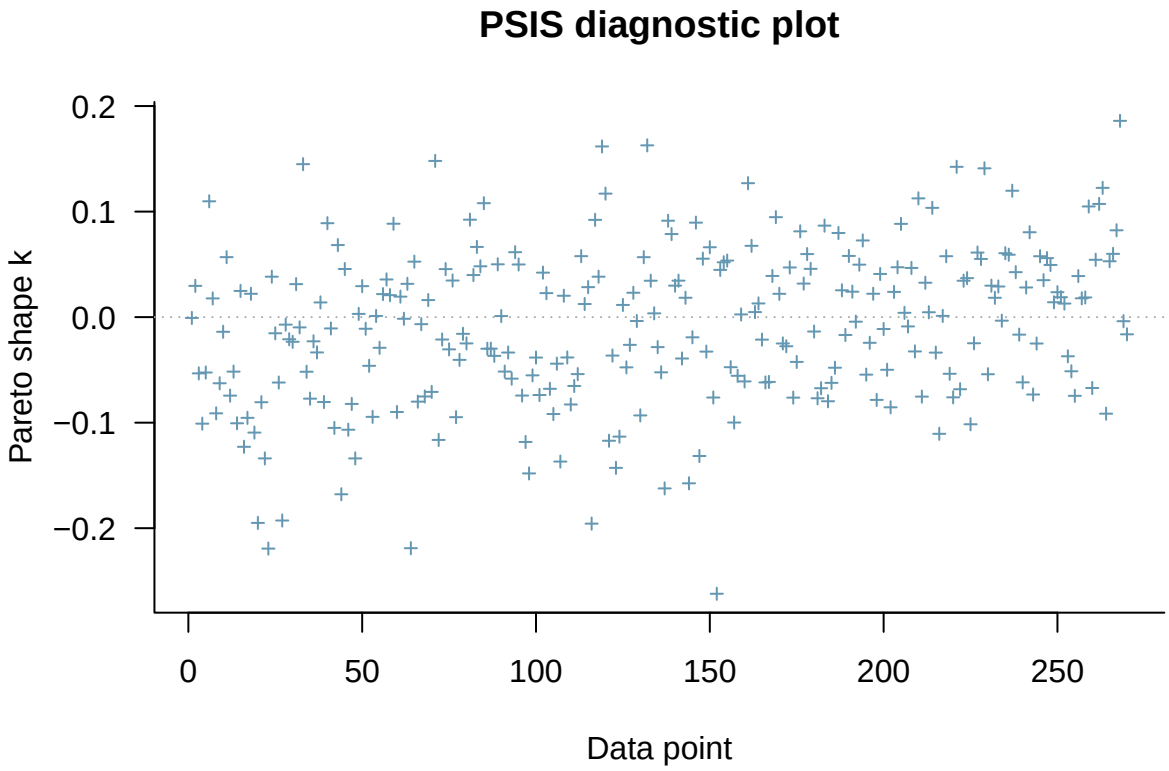


5 Evaluating the model

5.1 Using the `loo` function

To evaluate this model, we will utilize the `loo` function to look at different metrics. The expected log-predictive density (ELPD) is approximately -154.0, and the Pareto-k estimates are okay with values less than 0.5, indicating that the model is not sensitive to any observations in the dataset.

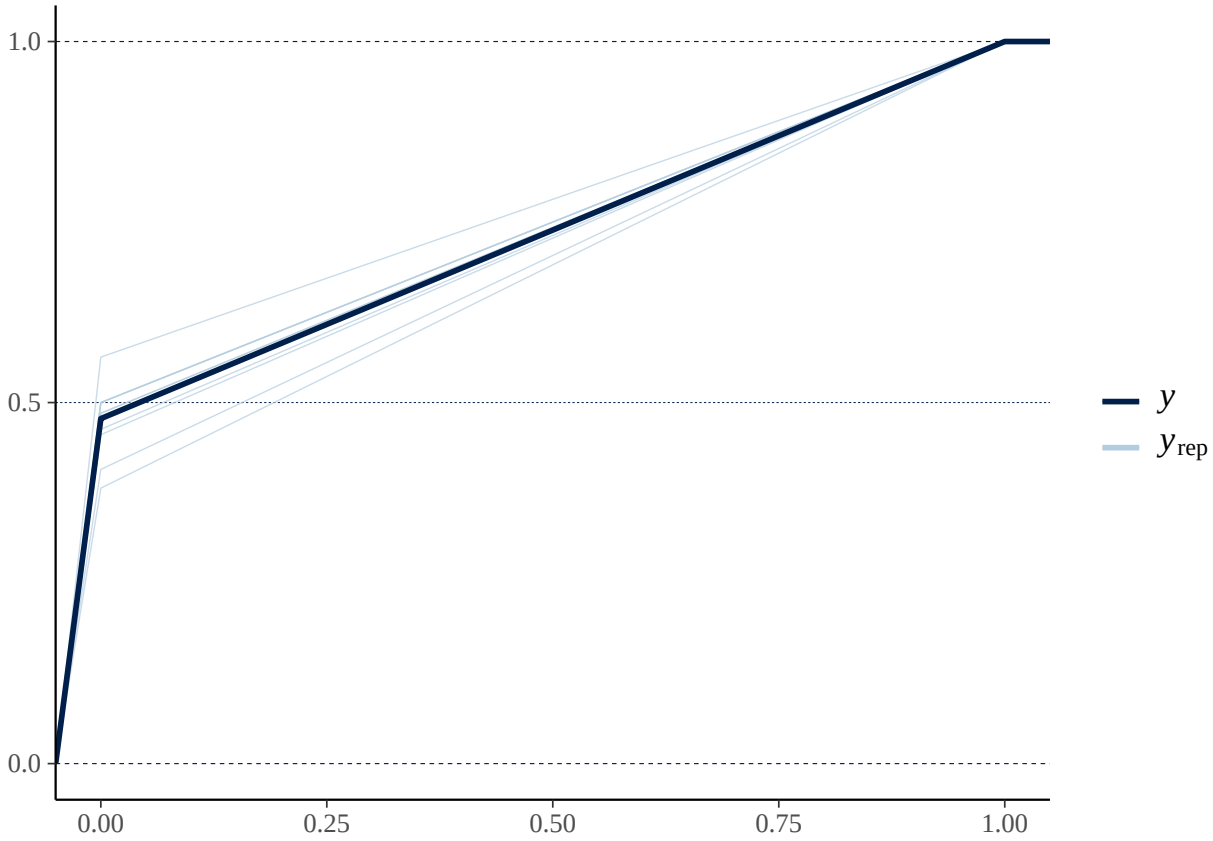
	Estimate	SE
elpd_loo	-154.03	7.10
p_loo	3.68	0.31
looic	308.05	14.21



5.2 Checking model fit

Next we will check the predicted values of the model compared to the actual training data. The ECDF plot below shows that the model is a decent fit to the data:

```
## Using 10 posterior samples for ppc type 'ecdf_overlay' by default.
```



5.3 Predictions on New Data

The model seemed to perform well on the training data and predict outcomes effectively. We will now use this model to predict playoff appearances for the 2019 season (our test set).

The code below shows that we will use our posterior distribution to predict playoff teams and then test accuracy:

```
nbatest$preds <- ifelse(colMeans(posterior_predict(post, newdata = nbatest)) > 0.5,1,0)
(sum(nbatest$Playoffs == nbatest$preds)/nrow(nbatest))
```

```
## [1] 0.6333333
```

The results show that the model does not do an excellent job at predicting playoff appearances on new data, with an accuracy of less than 70%.

6 Alternative Model

In the NBA, it is often preached that defense is the best offense, and that defense is the defining characteristic of a team that wins championships. The previous model gives some insight into what characteristics are common in playoff teams, but does not explain the data well, and misses a lot of variables that capture the intricacies of a team.

Let's extend this model to include additional variables. In this alternative model, we will include variables related to defensive characteristics of teams, to see if teams strong in defense have higher likelihoods of going to the playoffs. Similar to the previous model, we will use a bernoulli data generating process. However, the model will incorporate the following variables as predictors:

- `FG_percent: Total number of shots scored divided by total shots attempted.`
- `REB: Total rebounds per game.`
- `AST: Total assists per game.`
- `X3P_percent: Total number of 3-point shots scored divided by total 3-point shots attempted.`
- `DR: Total defensive rebounds per game.`
- `FTM: Total free-throws made per game.`
- `TO: Total turnovers made per game.`
- `STL: Total steals per game.`

The model will now take on the form:

$$logit(P_{Playoffs}) = \beta_0 + \beta_1 FGpercent + \beta_2 REB + \beta_3 AST + \beta_4 3Ppercent + \beta_5 DR + \beta_6 FTM + \beta_7 TO + \beta_8 STL$$

We will fit the following priors shown in the code below, and draw from the posterior distribution of the alternative model:

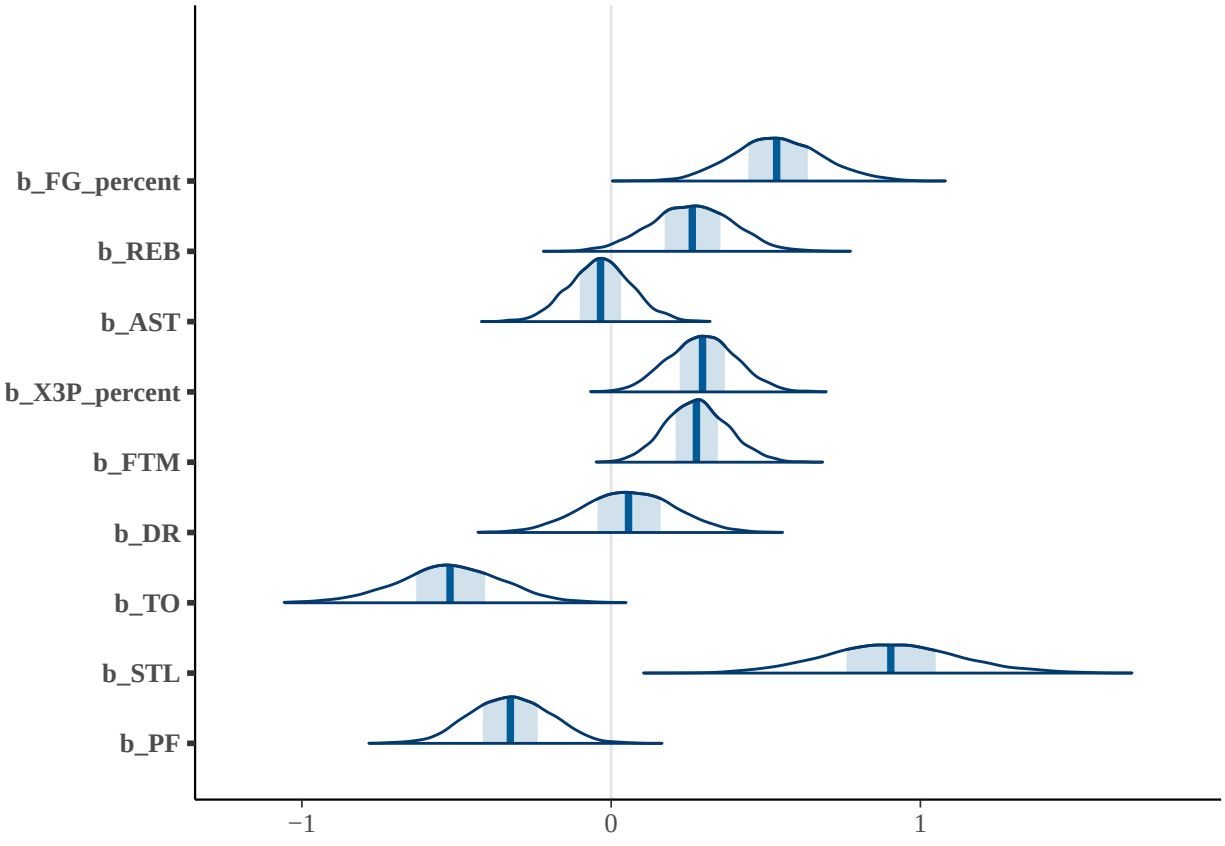
```
priors <- priors <- prior(normal(0, 1), class = "Intercept") +
  prior(normal(.20, 0.40), coef = "FG_percent") +
  prior(normal(0.20, 0.40), coef = "X3P_percent") +
  prior(normal(0.20, 0.20), coef = "AST") +
  prior(normal(0, 0.40), coef = "REB")
  prior(normal(0.20, 0.30), coef = "FTM") +
  prior(normal(0.10, 0.20), coef = "DR") +
  prior(normal(-0.50, 0.30), coef = "TO") +
  prior(normal(0.30, 0.30), coef = "STL")

post_alt <- brm(Playoffs ~ FG_percent + REB + AST + X3P_percent + FTM + DR + TO + STL + PF,
                data = nbatrain, family = bernoulli(), prior = priors)
```

6.1 Alternative Model Results

The result of the model is interesting, and seemingly confirms the theory that defense makes a team better. The coefficients on defensive variables `DR`, `STL`, `TO`, and `PF` all show much better likelihoods of being in the playoffs when those respective stats are better. For example, the estimated coefficient on `TO` is `-0.52`, meaning that teams with very high turnovers per game have a much lower chance of being a playoff team, because having high turnovers is indicative of a team with poor ball-handling ability.

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	-45.06	7.88	-61.28	-30.17	1	4475	3452
FG_percent	0.54	0.14	0.27	0.83	1	3664	3033
REB	0.26	0.13	0.01	0.51	1	2608	3112
AST	-0.03	0.10	-0.23	0.17	1	3579	2609
X3P_percent	0.30	0.11	0.09	0.51	1	3686	2964
FTM	0.28	0.10	0.09	0.48	1	3947	3486
DR	0.06	0.15	-0.23	0.34	1	2584	2761
TO	-0.52	0.17	-0.87	-0.20	1	4221	3352
STL	0.91	0.22	0.49	1.35	1	4022	3012
PF	-0.33	0.13	-0.57	-0.08	1	4193	2816

The plot below shows the updated posterior distribution of the impact on log-odds of playoff appearance by each predictor. It is apparent that defensive characteristics have a distribution corresponding to increasing log-odds.



6.2 Comparing to previous model

Compared to the previous model, the ELPD is higher, indicating that it's a better model fit. Also, like the previous model, all Pareto-k estimates are okay with values less than 0.5.

	Estimate	SE
elpd_loo	-138.89	9.21
p_loo	10.35	1.05
looic	277.78	18.42

	elpd_diff	se_diff	elpd_loo	se_elpd_loo	p_loo	se_p_loo	looic	se_looic
post_alt	0.00	0.00	-138.89	9.21	10.35	1.05	277.78	18.42
post	-15.14	6.95	-154.03	7.10	3.68	0.31	308.05	14.21

Evaluating the alternative model on the test data yields slightly better results than the previous model. The model predicts the out-of-sample test data with 73% accuracy.

```
nbatest$preds_alt  <- ifelse(colMeans(posterior_predict(post_alt, newdata = nbatest))>0.5,1,0)
sum(nbatest$Playoffs == nbatest$preds_alt)/nrow(nbatest)
```

```
## [1] 0.7333333
```

7 Conclusion

Comparing both models, it is clear the alternative model with more predictors is the better model with more accurate predictions. Overall, the performance is decent, and the model could be used for practical purposes to determine if a team is likely to be a playoff team during the course of the season.