A Comparative Study of Bayesian and Frequentist Logistic Regression for Army Recruitment Modeling Using Evidence from NCAA March Madness Data

Blake Coston

May 2025

1 Data

This project utilized two sources of data on NCAA mens March Madness tournaments from the years 2013 to 2025, both pulled from the Kaggle website. The first set contained variables about each team for each given year while the second had the layout of each tournament[1].

Our data had a 3 variables missing about 90% of the observations, so we removed them. Other than those three, only EFG_o and EFG_d had missing observations with about 7% and 15% missing respectively. We imputed the mean of these columns for those observations.

We joined the two sets of datasets into one data frame with the observations being a single game. We had a variable for each team name, each team seed, the winner, if the result was an upset, the round, and the difference between the other variables for each team (higher seed - lower seed). We chose to take the difference because it reduced dimensionality while retaining explainability. If not, we would have had almost double the number of variables.

[1] Barttorvik.com. *2025–26 Projections – Customizable College Basketball Tempo Free Stats – T-Rank*. `https://www.barttorvik.com/trankpre.php`. Accessed 2 Mar. 2025. 2025.

Variable	Description
upset	(Target Variable)
round	Round of the playoff 1-6
wins	Difference in wins
year	Tournament year
seed	Difference in seed
games	Difference in games played in the season
win_prop	Difference in win proportion
fg_for	Difference in field goal percentage for
fg_allowed	Difference in field goal percentage allowed
turn_for	Difference in turnover percentage for
turn_allowed	Difference in turnover percentage allowed
reb_for	Difference in rebound percentage for
reb_allowed	Difference in rebound percentage allowed
free_for	Difference in free throw percentage for
free_allowed	Difference in free throw percentage against
two_for	Difference in two point percentage for
two_allowed	Difference in two point percentage allowed
three_for	Difference in three point percentage for
three_allowed	Difference in three point percentage allowed
wins_above_bubble	Difference in number of wins over a team in the tournament
eff_o	Difference in offensive efficiency against an average D1 team
eff_d	Difference in defensive efficiency against an average D1 team
bar	Difference in bar score

Table 1: Variable Descriptions

1.1 Data Exploration

When looking at the histograms of the independent variables, we see that they do not have significant outliers or heavy tails. This is important because the less variance in the distributions, the closer the estimates of the parameters will be to the true parameters.

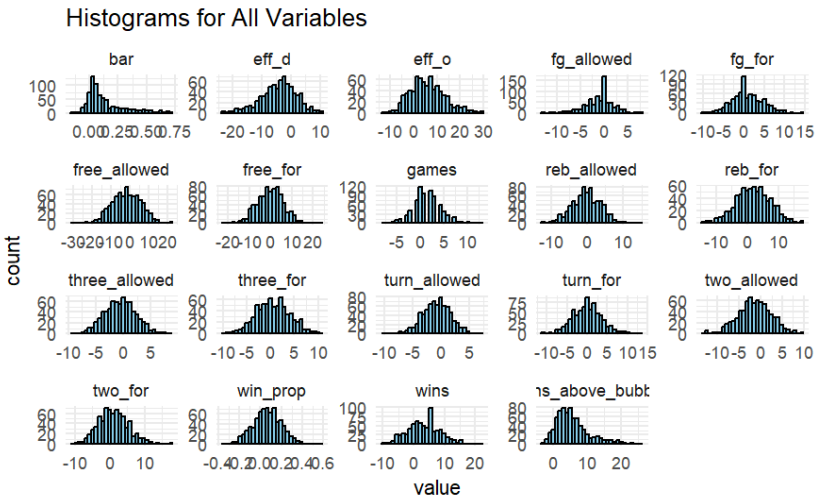

Figure 1: Histogram of Variables

Next, we found extremely high covariation between many of our variables. Appendix 2 displays the variables with variation outside the range $[-.4, 4]$. Some variables with high frequency of correlation between eachother and the other variables are: bar, eff_o, eff_d, and wins_above_bubble. These variables could cause multicollinearity in our model making estimates volatile and less accurate interpretations. They could also explain much of the variation in many of the other variables removing the need for those variables.

2 Methodology

2.1 Frequentist Logistic Regression

To model probabilities with frequentist statistical methods, we used logistic regression. Logistic regression is a form of generalized linear models (GLM) that models the probability of a binary outcome. A GLM consists of a linear predictor and a link function.

2.1.1 Linear Predictor

The linear predictor in any GLM is a function that linearly combines the estimated parameters and variables. The outcome η is not often an estimate of the response variable and is mapped to the mean of the response distribution by the inverse of the link function discussed in the next section[2]. In equation one, the β are parameters our model estimates, while x_n are the variables, and η is the response.

[2] Alan Agresti. *Foundations of Linear and Generalized Linear Models.* John Wiley & Sons, 2015.

Linear Predictor

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon_i = X^T \beta \tag{1}$$

In our logistic regression model, the linear predictor uses maximum likelihood estimation (MLE) to estimate the model parameters. The likelihood function used in MLE is the joint probability distribution of the response which is a function of the model parameters that calculates the probability of the observed data. As the name suggests, MLE seeks to maximize this function to find the best estimates of the parameter values.

Our likelihood function uses the sigmoid function to map the response η from the linear predictor to a value between 1 and 0 following the Bernoulli distribution, which models the probability of a single event.

Sigmoid Function

$$\sigma(\eta) = \frac{1}{1 + e^{-\eta}} \tag{2}$$

We then assume that the probability of a "success" p for a single event $y_i \in 0, 1$ follows a Bernoulli distribution.

$$y_i \sim Bernoulli(p) \tag{3}$$

Bernoulli Distribution

$$f(p) = p^{y_i}(1-p)^{1-y_i} \tag{4}$$

By replacing p with the sigmoid function of the linear predictor for observation i, $\sigma(X_i^T \beta)$ we get the likelihood for a single observation.

Probability of Observation i

$$P(y_i = 1|X_i) = (\sigma(X_i^T \beta))^{y_i} \ (1 - \sigma(X_i^T \beta))^{1-y_i} \tag{5}$$

Because the observations are independently identically distributed (iid.), the joint probability density function/likelihood function is the product across all observations. Maximizing this function yields the parameter estimates β used in the linear predictor equation 1.

Likelihood Function

$$L(\beta) = \prod_{i=1}^{n} (\sigma(X_i^T \beta))^{y_i} (1 - \sigma(X_i^T \beta))^{1-y_i} \tag{6}$$

2.1.2 Link Function

In GLMs, the link function translates the desired outcome, probability in this case, to the outcome η of the linear predictor. The inverse of the link function maps the linear predictor to the distribution of the response. All models in this project use the logit link function with the sigmoid function as the inverse.

Link Function

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \eta \tag{7}$$

The logit function equates the linear predictor to the log odds which can then be further transformed into the probability of an event using the sigmoid function. To use this model to predict, we map the probability to $1, 0$ using a cut-off point for a "success".

Probability Calculation

$$p = \frac{1}{1+e^{-\eta}} \tag{8}$$

2.2 Bayesian Logistic Regression

Bayesian and frequentist logistic regression models estimate the same log odds but use different methods. Bayesian methods are fundamentally built on bayes theorem which is used in bayesian regression to update our prior belief on a subject with data.

Bayes Theorem

$$P(A|B) = \frac{P(A)(B|A)}{P(B)} \tag{9}$$

Instead of point estimates, like in frequentist statistics, bayesian statistics treats parameters as random variables and estimates the distribution of the parameters conditional on the data and factoring in prior belief. The resulting distribution is called the posterior distribution. To determine this posterior distribution, we use the basic formula

Bayesian Formula

$$posterior \propto prior \times likelihood \tag{10}$$

This formula states that the posterior joint probability of the parameters is proportional to the prior distributions times the likelihood given the data. In the formula for equation 10, the prior distribution of the parameters can be well informed, uninformed, based on previous data, or just our prior knowledge on the subject, but must be independent from the data used in the likelihood. The joint prior distribution is the product of the individual prior distributions. The likelihood is the same likelihood function as the one used in the frequentist MLE. By multiplying the priors by the likelihood function, we get the shape of the posterior. We then need to divide this by a constant so that the area under the curve equals one, creating the posterior a probability distribution. This constant, called the marginal likelihood, integrates the likelihood function times the priors over the parameter space and acts as a normalization or scaling factor for the posterior so that it is a probability density function.

Like in logistic regression, we use the Bernoulli distribution to model the probability p of a "success" given the data y. In bayesian logistic regression this is the form of the marginal posterior.

Posterior Distribution Function

$$f(\theta|y) = \frac{g(\theta) \times L(y|\theta)}{\int g(\theta) \times L(y|\theta) d\theta} \tag{11}$$

In the equation, $f(\theta|y)$ is the joint distribution of the parameters θ conditional on the data y, $g(\theta)$ the joint prior distributions, $L(y|\theta)$ the likelihood function, and $\int g(\theta) \times L(y|\theta) d\theta$ the marginal likelihood or normalization factorWilliam M. Bolstad. *Introduction to Bayesian Statistics*. John Wiley & Sons, 2013. To make predictions, we can use the mean of the marginal probability distributions of the parameters as point estimates similar to frequentist logistic regression.

Bayesian models use one of many sampling methods to estimate parameter distributions while considering the preset priors. The power of Bayesian methods comes from the ability to include prior belief in the model instead of relying solely on the data. This is especially useful when we are well informed on the subject we are modeling.

3 Models

3.1 Overview

This section outlines the different models used on the March Madness dataset. We fit five logistic regression models and four bayesian logistic regression models on a train dataset then compared them with a test dataset and the 2025 bracket. The goal of these models is to most accurately predict upsets to provide the best interpretation of the variables. However, because we are only interested in how these models would perform on Army recruiting data, we will only analyze a few variables and their effects on the probability of an upset to show the process. We will not go in depth into basketball statistics or make conclusions about the tournament.

3.2 Frequentist Models

We fit five logistic regression models on our March Madness dataset using different variable selection methods. The first model included all variables listed in Table 1. In this model, only seven out of 22 variables were significant. Also, because many of the variables were significantly correlated which makes it hard to determine the effects from the individual variables on the response known as multicollinearity.

To combat issues with the first model, we created a second model which reduced the number of parameters using a stepwise function. The new model used bidirectional stepwise selection, minimizing AIC, and produced a model with variables: wins, win_prop, turn_for, reb_allowed, free_allowed, two_allowed, three_allowed, wins_above_bubble, eff_o, and eff_d.

Variable	**Estimate**	**Std. Error**	**Significance**
(Intercept)	-0.221	0.196	
wins	-0.655	0.079	***
win_prop	7.733	2.972	**
fg_allowed	-0.512	0.108	***
turn_for	0.408	0.096	***
reb_allowed	-0.322	0.064	***
free_allowed	-0.041	0.024	.
wins_above_bubble	0.583	0.087	***
eff_o	-0.225	0.035	***
eff_d	0.589	0.085	***

Table 2: Stepwise Logistic Regression Results

Though this model minimized AIC, it still contained many significantly correlated pairs, which could make the estimates volatile. This volatility is evident in the change in parameter estimates in many of the variables. For example, the parameter for win_prop in our full logistic model had an estimate of 3.787, but

in our stepwise model the estimate was 7.82. This significant change in estimate drastically throws off our interpretation of the parameter.

So, we created a third model that removed variables found in many significant pairs from Appendix 2. This model did not include the variables: two_for, two_allowed, wins_above_bubble, eff_o, eff_d, bar. We then performed the same stepwise regression on that model to reduce the number of variables. This second step model included only five variables, of which only two were significant meaning we cannot make many conclusions about the model.

Variable	**Estimate**	**Std. Error**	**Signif.**
(Intercept)	43.59	66.72	
round	0.015	0.106	
wins	-0.084	0.228	
year	-0.022	0.033	
seed	-0.037	0.037	
games	-0.463	0.187	*
win_prop	1.809	7.723	
fg_for	-0.063	0.056	
fg_allowed	0.036	0.056	
turn_for	-0.053	0.050	
turn_allowed	0.033	0.064	
reb_for	-0.030	0.028	
reb_allowed	-0.024	0.036	
free_for	0.022	0.023	
free_allowed	0.000	0.021	
three_for	-0.043	0.049	
three_allowed	0.045	0.053	

Table 3: Uncorrelated Model

The final model we created includes variables found significant in another study which calculated variables in a similar manner (add link): eff_o, eff_d, bar, and two_allowed[3].

Variable	**Estimate**	**Std. Error**	**Signif.**
(Intercept)	0.211	0.145	
eff_o	-0.211	0.042	***
eff_d	0.261	0.052	***
bar	5.927	2.763	*
two_allowed	-0.062	0.036	.

Table 4: Archiv Model

[3]Christian McIver, Karla Avalos, and Nikhil Shivakumar Nayak. *March Madness Tournament Predictions Model: A Mathematical Modeling Approach.* 7 pages, 5 figures. 2025. arXiv: 2503.21790 [stat.AP]. URL: https://doi.org/10.48550/arXiv.2503.21790.

From these models, we selected the stepwise model variables for the bayesian models as it is simple while maintaining accuracy and precision as discussed in the model selection section.

3.3 Bayesian Models

Using the variables selected in the first stepwise logistic regression model, we fit four Bayesian models. The first model has uninformed normal priors with a mean of zero and a standard deviation of one. The mean of the parameter distributions for this model were almost exactly the same as the estimates from the logistic regression model with the same variables. This is as expected because the priors were uninformed, so the distributions were only based on the data.

Variable	**Mean**	**SD**	**10%**	**50%**	**90%**
(Intercept)	-0.2	0.2	-0.4	-0.2	0.0
wins	-0.6	0.1	-0.7	-0.6	-0.5
win_prop	3.3	1.9	0.9	3.3	5.7
fg_allowed	-0.6	0.1	-0.7	-0.6	-0.5
turn_for	0.5	0.1	0.4	0.5	0.6
reb_allowed	-0.4	0.1	-0.5	-0.4	-0.3
free_allowed	-0.1	0.0	-0.1	-0.1	0.0
wins_above_bubble	0.6	0.1	0.5	0.6	0.7
eff_o	-0.2	0.0	-0.3	-0.2	-0.2
eff_d	0.7	0.1	0.6	0.7	0.8

Table 5: Bayesian Logistic Regression Results

To improve upon this model with informed priors, we included priors found from a similar study conducted by members of the Pomona College Economics Department (2012)[4]. This study used a probit model instead of logit and used both opponent and team stats instead of taking the difference between the variables. So, we converted the probit variables into their logit counterparts and then took the difference between them, opponent - team. This results in the effect of the difference in variables on the log odds that the opponent, lower seed in our case, wins. We then converted the standard error to standard deviation and found the joint standard deviation. We used these informed mean and standard deviations for the normal distributions of variables: win_prop, fg_for, eff_o, and eff_d. For the other variables, we kept the uninformed normal priors at a mean of zero and a standard deviation of one.

The third was fit using the first year in the dataset, then used those parameter distributions as priors for the next year and so on until a model was fit for the 2024 season. We created this model attempting to bias later years

[4]Ross Steinberg and Zane Latif. "March Madness Model Meta-Analysis: What Determines Success in a March Madness Model?" Senior Paper in Economics, Pomona College, supervised by Professor Gary Smith. 2018.

over earlier ones. For our fourth model, to further bias more recent years, we used the same year-based model but increased the standard deviations of the priors by a factor of two. This second model did not improve upon the first as discussed in the next section.

Variable	Mean	SD	10%	50%	90%
(Intercept)	-0.4	0.5	-1.0	-0.4	0.2
wins	-0.8	0.0	-0.9	-0.8	-0.7
win_prop	4.4	1.3	2.7	4.4	6.1
fg_allowed	-0.9	0.1	-1.0	-0.9	-0.8
turn_for	0.8	0.0	0.7	0.8	0.8
reb_allowed	-0.5	0.0	-0.6	-0.5	-0.5
free_allowed	-0.1	0.0	-0.1	-0.1	0.0
wins_above_bubble	1.0	0.0	0.9	1.0	1.1
eff_o	-0.3	0.0	-0.4	-0.3	-0.3
eff_d	1.0	0.0	0.9	1.0	1.0

Table 6: Bayesian Year Priors

3.4 Model Selection

When selecting the best model, we looked at the confusion matrix, precision, accuracy, recall, specificity, and f1 score for each model based on our predictions on both the test set and the 2025 March Madness bracket. The test set consisted of 134 observations, 45 of which were upsets and 89 were not. This yields an accuracy of 66% using the dummy variable not an upset and predicting the higher seed to win every game.

Name	TP	TN	FP	FN
logistic_model	25	81	8	20
step_model	23	81	8	22
uncor_model	26	83	6	19
step2_model	24	80	9	21
archiv_model	9	81	8	36
bayes_model	25	80	9	20
prior_bayes_model	24	80	9	21
bayes_years_model	27	80	9	18
bayes_years_flat_model	26	81	8	19

Table 7: Test Set Confusion Matrix

Table 2 shows the confusion matrix for each of the models on the test set. All of the models performed similarly when predicting true negatives. Other than the archiv_model, which performed drastically worse than the other models, the true positive rate was relatively similar. Many of the models have a high number

of false negatives and low number of false positives meaning they under predict upsets.

Name	Accuracy	Precision	Recall	Specificity	F1
logistic_model	0.7910	0.7576	0.5556	0.9101	0.6410
step_model	0.7761	0.7419	0.5111	0.9101	0.6053
uncor_model	0.8134	0.8125	0.5778	0.9326	0.6753
step2_model	0.7761	0.7273	0.5333	0.8989	0.6154
archiv_model	0.6716	0.5294	0.2000	0.9101	0.2903
bayes_model	0.7836	0.7353	0.5556	0.8989	0.6329
prior_bayes_model	0.7761	0.7273	0.5333	0.8989	0.6154
bayes_years_model	0.7985	0.7500	0.6000	0.8989	0.6667
bayes_years_flat_model	0.7985	0.7647	0.5778	0.9101	0.6582

Table 8: Performance Metrics on Test Set

Table 3 shows the performance metrics for each of the models on the test set. First, accuracy, we can see that all models significantly improved on the dummy model that just predicted the higher seed would win. Next, precision shows the percent of predicted upsets that were actually true. Again, all models but the archiv_model performed simlarly with the archiv_model being significantly worse. Recall shows the proportion of actual positives that were correctly identified while specificity shows the proportion of actual negatives that were correctly identified. All models but the archiv_model correctly identified slightly over half the upsets, but all models correctly predicted about 90% of the non upsets. Finally, f1 score is used to balance precision and recall through the harmonic mean.

While these metrics are good at comparing within the time frame of the dataset, we want to extrapolate to future years. So, we tested the models on the 2025 bracket. Because the models may not predict the correct winner from previous rounds, we cannot use the same metrics as we did on the test dataset. Instead, we used the ESPN bracket scoring. This scoring method is out of 1920 points. For each correctly predicted winner, a bracket is awarded points for each round. Table 4 shows a breakdown of the scoring process while Table 5 shows the score for each of the models.

Round Name (number)	Correct Pick Points
Round of 64 (1)	10
Round of 32 (2)	20
Sweet 16 (3)	40
Elite 8 (4)	80
Final Four (5)	160
Championship Game (6)	320

Table 9: ESPN Scoring System

Model	ESPN Points	Percentage Correct
logistic_model	1010	63.49%
step_model	1140	68.25%
uncor_model	1200	60.32%
step2_model	1100	65.08%
archiv_model	1280	80.95%
bayes_model	1220	73.02%
prior_bayes_model	1100	66.67%
bayes_years_model	700	58.73%
bayes_years_flat	710	60.32%
Seed Based Model	1250	77.78%

Table 10: ESPN Points Comparison

The 2025 bracket was unique in that the seed based model performed significantly better than the other models. Where in our test set we had 34% upsets, the 2025 bracket only had 23%. This drastically favors models like the archiv_model that are more conservative in selecting an upset.

So, our recommendations based on the test dataset and the 2025 bracket are to keep the uncor_model, archiv_model, and bayes_years_model. The uncor_model performed drastically better on the test dataset than the other models and was one of the better models on the 2025 dataset. The archiv_model is extremely conservative, so it performed poorly on the test dataset, but when extrapolating it to the 2025 tournament it was the only model to do better than the seed based model. Finally, the bayes_years_model performed just behind the uncor_model, but with the variables from the first step model which it drastically outperformed. With the variables from the uncor_model, we could create a bayesian years based model that performs better than the uncor_model.

References

Agresti, Alan. *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons, 2015.

Barttorvik.com. *2025–26 Projections – Customizable College Basketball Tempo Free Stats – T-Rank*. https://www.barttorvik.com/trankpre.php. Accessed 2 Mar. 2025. 2025.

Bolstad, William M. *Introduction to Bayesian Statistics*. John Wiley & Sons, 2013.

McIver, Christian, Karla Avalos, and Nikhil Shivakumar Nayak. *March Madness Tournament Predictions Model: A Mathematical Modeling Approach*. 7 pages, 5 figures. 2025. arXiv: 2503.21790 [stat.AP]. URL: https://doi.org/10.48550/arXiv.2503.21790.

Steinberg, Ross and Zane Latif. "March Madness Model Meta-Analysis: What Determines Success in a March Madness Model?" Senior Paper in Economics, Pomona College, supervised by Professor Gary Smith. 2018.

4 Appendices

Appendix 1: Variable Statistics

Variable	Min	Max	Mean
round	1	6	1.85
year	2013	2025	NA
team1_seed	1	16	4.84
team2_seed	1	16	8.94
upset	0	1	0.29
seed	1	15	6.53
games	-8	13	1.41
win_prop	-0.35	0.60	0.07
fg_for	-10.2	14.0	0.89
fg_allowed	-13.1	7.6	-1.03
turn_for	-11.1	13.4	-0.00
turn_allowed	-10.4	7.0	-0.63
reb_for	-14.1	17.5	1.96
reb_allowed	-12.6	14.6	-0.22
free_for	-22.4	23.8	-0.12
free_allowed	-29.9	22.3	-1.80
two_for	-10.3	17.9	1.08
two_allowed	-13.3	9.5	-1.42
three_for	-9.7	11.0	0.57
three_allowed	-9.4	8.5	-0.57
wins_above_bubble	-3.2	25.7	5.65
eff_o	-13.1	29.1	6.00
eff_d	-23.4	10.6	-4.15
bar	-0.16	0.53	0.12
wins	-10	22	3.39

Appendix 2: Top Variable Pair Correlations

Var1	Var2	Correlation
wins	win_prop	0.8913
wins_above_bubble	seed	0.8546
bar	wins_above_bubble	0.8539
bar	seed	0.8442
two_for	fg_for	0.8288
two_allowed	fg_allowed	0.7923
bar	eff_o	0.7072
three_for	fg_for	0.6997
bar	eff_d	-0.6793
eff_o	wins_above_bubble	0.6528
eff_o	seed	0.6286
wins	wins_above_bubble	0.6213
eff_d	fg_allowed	0.6142
eff_d	wins_above_bubble	-0.6030
eff_o	fg_for	0.5819
eff_d	seed	-0.5780
eff_d	two_allowed	0.5682
three_allowed	fg_allowed	0.5469
free_allowed	turn_for	0.5336
eff_o	two_for	0.5194
wins_above_bubble	win_prop	0.5150
wins	games	0.4984
games	upset	-0.4818
eff_o	three_for	0.4713
reb_allowed	turn_for	0.4350
eff_d	three_allowed	0.4198
fg_for	win_prop	0.4149
eff_o	turn_allowed	-0.4117
wins	bar	0.4094
two_for	win_prop	0.4032