# Bayesian March Madness Model Design

## 1. Data Collection (2021–2024)

Collected historical NCAA basketball tournament data from 2021 to 2024 using Barttorvik and KenPom databases. Scraped advanced metrics including Adjusted Offensive Efficiency (AdjOE), Adjusted Defensive Efficiency (AdjDE), Barthag ratings, effective FG%, turnover rates, rebound rates, free-throw rates, and shooting percentages.

## 2. Feature Selection and Composite Strength Index

Created a 'Composite Strength' index calculated as the average of standardized values from the following metrics:

* **Overall Strength:** Bart Torvik's “Barthag” aggregate team strength metric, incorporating “wins above bubble (WAB)”
* **Barthag logit:** Log-odds transformation of **Barthag rating**, basically a KenPom **expected win percentage** converted to log-odds.
* **AdjOE:** Adjusted Offensive Efficiency (KenPom).
* **AdjDE:** Adjusted Defensive Efficiency (KenPom).
* **Clutch Index:** Team performance in high-pressure situations (close games), weighted by a one-hot encoding of whether they made it to each level of the tournament (Round of 32, sweet 16, elite 8, etc)
* **Upset Factor:** Team's propensity for tournament upsets based on historical seed-based performance.
* **Turnover Edge:** Differential between opponent (TORD) and team (TOR) turnover rates.
* **Conference Strength:** Average number of teams within the same conference that made it to R32. In the data, R32 is 1 or 0 if they team made it. (Definitely could make this one better)

## 3. Explicitly Derived Metrics

Defined metrics explicitly as follows:

* Upset Factor =
* Turnover Edge =
* Clutch Index =
* Barthag Logit =
* Overall Strength =
* Conference strength =

## 4. Bayesian Logistic Regression Model

This model uses historical tournament data (overall team strength, efficiency ratings, clutch performance, conference strength, and upset likelihood) from the 2021-2022 to 2023-2024 seasons to estimate the log-odds of winning the NCAA championship. It uses a Bayesian logistic regression model with a mixed-effects structure:

**Priors:**  
- Intercept (β₀): - Coefficients (β₁–β₈):   
- Conference random intercepts ():

(Note: These priors could be changed. I made the intercept weakly informative, but I made the coefficients prior strongly regularizing, which means that mainly strong effects will persist while “noise” is filtered out. I didn’t think about this too closely, probably warrants another look.)

I chose a hierarchical structure because I felt that conference-level differences would be expected to systematically influence championship outcomes, and a random intercept allows these effects to be modeled while preventing overfitting. The random intercept accounts for this conference-level variability, allowing teams from stronger conferences to have different baseline probabilities while still sharing statistical information across conferences. Basically, in a hierarchical (mixed-effects) model, random effects (e.g., ) are modeled as being drawn from a shared normal distribution rather than estimated as fixed parameters. So this parameter is directly estimated from the data (conferences with little data are shrunk toward the overall mean rather than being overfitted, and information from conferences that systematically produce stronger teams is learned and incorporated).

Then the model is fit via Hamiltonian Monte Carlo (4 chains, 4000 iterations each), which is pretty standard for this type of model. Once we fit this model, we obtain **posterior distributions** for all regression coefficients. These posteriors quantify the uncertainty in how each feature (e.g., offensive efficiency, clutch index) influences championship probability. They also serve as Bayesian priors for future simulations.

The posterior distributions serve as Bayesian priors for future simulations specifically in two key ways:

1. **When computing each team’s expected strength:** The model draws β coefficients (e.g., β₁, β₂, … β₈) from the posterior distributions, meaning that each feature’s contribution to team strength reflects the uncertainty learned from historical data.
2. **When incorporating conference-level effects:** The random intercepts () are drawn from their posterior distributions, ensuring that the prior information about conference strength carries forward into new season simulations.

## 5. Probabilistic Matchup Simulations

Now, instead of directly predicting the overall champion, we use these posterior distributions to simulate individual game matchups using the current season’s data (2024-2025). Each team’s expected strength is calculated using Bayesian posterior draws, incorporating the priors learned from historical data.

1. The **priors for β coefficients** (e.g., β₁, β₂, … β₈) are used here, meaning each team’s strength is modeled **with** **uncertainty** based on past seasons.
2. **The priors for conference effects** () ensure that teams from historically strong conferences retain some advantage, but with **updated information from the current season**.

This allows us to estimate a linear predictor for each team:

This as a Bayesian-simulated score that reflects the team’s predicted strength based on their features. We then compare LP values between two teams to estimate win probabilities using a logistic transformation.

For a given matchup, the win probability is calculated:

This difference quantifies how much stronger Team A is expected to be compared to Team B, integrating Bayesian priors at every step to ensure predictions reflect both historical data and current season performance while incorporating uncertainty. The logistic (plogis, or sigmoid) function then converts these simulated Bayesian linear predictors into a win probability for Team A, constraining the result between 0 and 1. Currently, the algorithm **picks a winner deterministically** (i.e., if , then Team A wins; if not, then Team B wins). Have been thinking about making this stochastic, but I’m not sure how to best implement that yet.

## 6. Multiple Bracket Simulations

Generated multiple bracket simulations by varying random seeds. Each simulation reflects the uncertainty captured by Bayesian posterior distributions, providing diverse bracket predictions.

## 7. Diagnostic Evaluation

Validated and visualized results through composite strength rankings, matchup probability distributions, and credible intervals, ensuring robust interpretation and clarity 🡪 Rmarkdown HTML document that you’ve seen.