Predicting 2025 NFL Wild Card Outcomes: A Hierarchical Bayesian MCMC Approach

Udayveer Singh Andotra (ua118)

960:568 Bayesian Analysis Spring 2025 Final Project Report

8th May 2025

Abstract

This project develops a hierarchical Bayesian Markov Chain Monte Carlo (MCMC) model to predict point differentials and win/loss outcomes for the 2025 NFL Wild Card games using regular season data from `NFL-DataScrap(2024-2025).csv`. The model uses team strengths, home-field advantages, and factors like `spread-line`, `off-epa-per-play`, `cpoe`, `wpa`, and points scored in the last three games (home and away), with parameters estimated using a Gibbs sampler. Results show balanced outcomes, but extreme predictions suggest sensitivity to unscaled factors. The model correctly predicted 4 out of 6 games, achieving a 66.67% accuracy. Violin plots provide insights into prediction distributions. Challenges like team mismatches and factor selection are discussed, along with future improvements such as scaling factors and adding defensive metrics.

1 Introduction

The National Football League (NFL) Wild Card games are a key part of the postseason, deciding which teams advance to the Divisional Round. Predicting game outcomes, such as point differentials and win probabilities, is challenging due to varying team performance, home-field effects, and game-specific factors. This project builds a statistical model to predict the 2025 NFL Wild Card game outcomes for matchups: Buffalo Bills vs. Denver Broncos, Baltimore Ravens vs. Pittsburgh Steelers, Houston Texans vs. Los Angeles Chargers, Philadelphia Eagles vs. Green Bay Packers, Tampa Bay Buccaneers vs. Washington Commanders, and Los Angeles Rams vs. Minnesota Vikings.

The data comes from `NFL-DataScrap(2024-2025).csv`, which includes regular season statistics for the 2024-2025 NFL season, such as `spread-line`, `off-epa-per-play`, `cpoe`, `wpa`, `home_points_last_3`, `away_points_last_3`, and team identifiers. A hierarchical Bayesian MCMC model is used, incorporating team strengths, home-field advantages, and other factors to predict point differentials (y_g = home points − away points). A Gibbs sampler estimates the model parameters, generating predictions for win/loss outcomes and point differential distributions. This report covers the methodology, results, challenges, and future directions.

2 Methodology

2.1 Data Preprocessing

The dataset `NFL-DataScrap(2024-2025).csv` contains game statistics for the 2024-2025 NFL regular season. Preprocessing steps include:

- **Normalization**: Team names were converted to lowercase (e.g., `buf`, `la`) for consistency with CSV abbreviations.
- **Team Mapping**: A mapping of abbreviations to full names (e.g., `buf` to Buffalo Bills) was created and saved as `team-mapping.csv` for readable outputs.
- **Challenge Resolution**: Initial `NA` entries due to mismatches (e.g., `buffalo` vs. `buf`) were fixed by normalizing team names and correcting the Rams abbreviation (`lar` to `la`).

2.2 Model Structure

A hierarchical Bayesian MCMC model predicts the point differential $y_g = \text{home points} - \text{away points}$ for game g. The model is defined with the following components:

The likelihood of y_g follows a normal distribution:

$$y_g \sim \mathcal{N}(\mu_g, \sigma^2), \quad \mu_g = \alpha_h - \alpha_a + \beta_h + \text{cov-adj},$$

where α_h is the home team's strength, α_a is the away team's strength, β_h is the home-field advantage, and cov-adj $= \gamma \cdot$ `spread-line` $+ \delta_1 \cdot$ `off-epa-per-play-h` $+ \delta_2 \cdot$ `off-epa-per-play-a` $+ \eta \cdot$ `cpoe` $+ \theta \cdot$ `wpa` $+ \lambda_1 \cdot$ `home_points_last_3` $+ \lambda_2 \cdot$ `away_points_last_3`.

The model parameters have the following priors:

- Team strengths: $\alpha_i \sim \mathcal{N}(\phi\alpha_{i,\text{prev}}, 1)$, with $\phi = 0.9$, to account for performance trends over time.
- Home-field advantage: $\beta_i \sim \mathcal{N}(0, 4)$, reflecting a typical NFL home advantage of 2 to 3 points, with variance for venue effects like crowd or weather.
- Covariate coefficients: $\gamma, \delta_1, \delta_2, \eta, \theta, \lambda_1, \lambda_2 \sim \mathcal{N}(0, 4)$, handling scales of `spread-line` (0 to 15), `off-epa-per-play` (0 to 0.5), `cpoe` (0 to 1), `wpa` (typically -0.5 to 0.5), and `home_points_last_3` and `away_points_last_3` (0 to 150) while preventing overfitting.
- Residual variance: $\sigma^2 \sim \text{Inv-Gamma}(5, 5)$, a prior with a mean of 1.25, suitable for noisy sports data.

The mean μ_g is a derived quantity from α_i, β_i, and covariate coefficients, so it does not require a direct prior. The `home/away-win-streak-last-3` factor was removed to simplify the model, focusing on offensive, performance, and recent scoring metrics.

2.3 MCMC Gibbs Sampler

A Gibbs sampler is used to estimate the posterior distribution $p(\alpha, \beta, \text{covariates}, \sigma^2 \mid y)$ by sampling from the conditional distributions of each parameter. The process is as follows:

The joint posterior combines the likelihood and priors:

$$p(\alpha, \beta, \text{covariates}, \sigma^2 \mid y) \propto \prod_g \mathcal{N}(y_g \mid \alpha_h - \alpha_a + \beta_h + \text{cov-adj}, \sigma^2) \cdot p(\alpha) \cdot p(\beta) \cdot p(\text{covariates}) \cdot p(\sigma^2).$$

- **Sampling** α_i: The conditional for α_i uses the likelihood and prior $\alpha_i \sim \mathcal{N}(\phi\alpha_{i,\text{prev}}, 1)$. For games involving team i, residuals $y_g - (\alpha_h - \alpha_a + \beta_h + \text{cov-adj})$ are adjusted by σ^2, balancing data fit and time trends.
- **Sampling** β_i: The conditional for β_i combines the likelihood (for home games of team i) and prior $\beta_i \sim \mathcal{N}(0, 4)$. Residuals are scaled by $1/\sigma^2$, resulting in a normal distribution with variance based on the prior precision $1/4$.
- **Sampling Covariate Coefficients (e.g.,** $\gamma, \eta, \theta, \lambda_1, \lambda_2$**)**: For coefficients like γ (for `spread-line`), η (for `cpoe`), θ (for `wpa`), λ_1 (for `home_points_last_3`), and λ_2 (for `away_points_last_3`), the conditional is:

  $p(\gamma, \eta, \theta, \lambda_1, \lambda_2 \mid y, \alpha, \beta, \text{other covariates}, \sigma^2) \propto \prod_g \mathcal{N}(y_g \mid \alpha_h - \alpha_a + \beta_h + \gamma \cdot \texttt{spread-line}_g + \eta \cdot \texttt{cpoe}_g + \theta \cdot \texttt{wpa}_g + \lambda_1 \cdot \texttt{home\_points\_las}$

  This gives a normal distribution, with the mean adjusted by residuals weighted by the respective covariate values and variance combining data precision and prior precision $1/4$. The posterior means of these coefficients quantify the impact of each covariate on point differentials, aiding interpretation of factor importance.
- **Sampling** σ^2: The conditional for σ^2 is an inverse-gamma distribution, using residuals $\sum_g (y_g - \text{predicted}_g)^2/2$ and prior Inv-Gamma$(5, 5)$ parameters (shape $5 + n_{\text{games}}/2$, rate $5 +$ sum of squared residuals$/2$).
- **Iteration and Convergence**: The sampler runs for 5000 iterations, discarding the first 1000 as burn-in to ensure convergence. The samples are used to compute predictions and win probabilities.

Team-specific parameters α_i and β_i are not summarized as mean coefficients because they are numerous and context-dependent, contributing directly to game-specific predictions rather than global effects.

3 Results

3.1 Win/Loss Outcomes

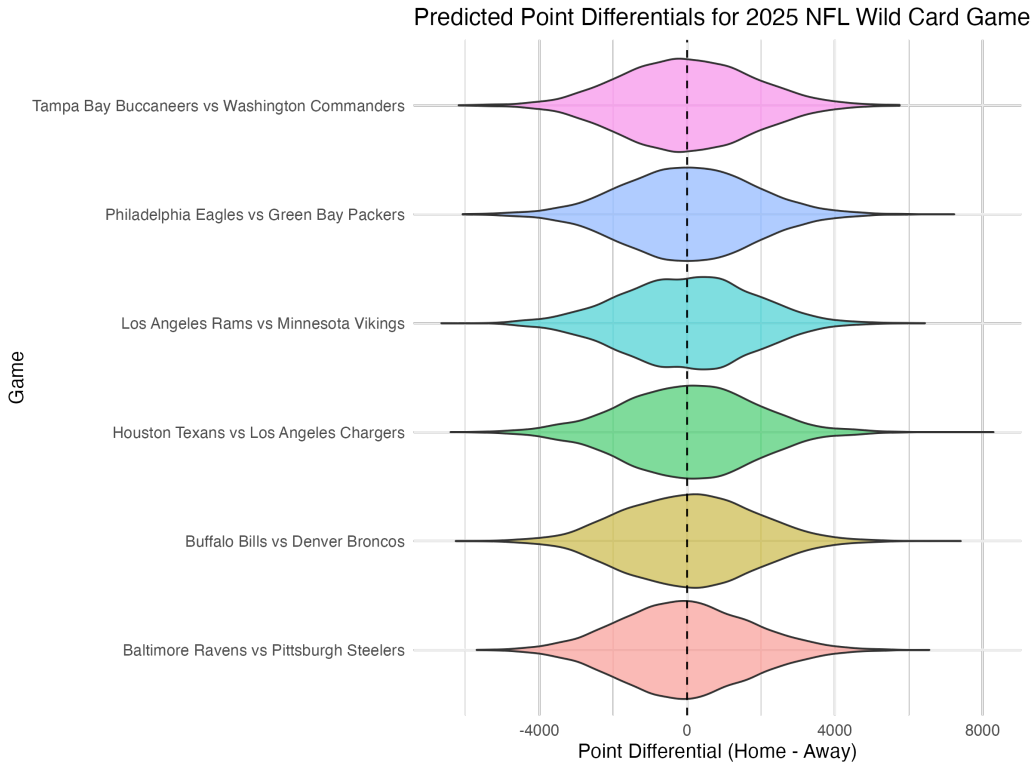
The model predicts point differentials and win/loss outcomes for the 2025 NFL Wild Card games using posterior samples. The table below shows the mean point differential, home win probability, and home team result for each matchup.

Game	Mean Point Differential	Home Win Probability	Home Team Result
Buffalo Bills vs Denver Broncos	46.40	0.51	Win
Baltimore Ravens vs Pittsburgh Steelers	-54.50	0.48	Loss
Houston Texans vs Los Angeles Chargers	29.70	0.51	Win
Philadelphia Eagles vs Green Bay Packers	24.70	0.51	Win
Tampa Bay Buccaneers vs Washington Commanders	-15.97	0.49	Loss
Los Angeles Rams vs Minnesota Vikings	-0.94	0.51	Loss

A positive mean point differential indicates a home team advantage, while the home win probability shows the proportion of samples favoring the home team.

3.2 Point Differential Distributions

The figure below shows a violin plot of the predicted point differential distributions, providing insights into prediction uncertainty.



Key observations:

- **Distributions**: The violin plot shows the spread of predicted point differentials for each game.
- **Extreme Spreads**: Wide distributions for Bills vs. Broncos (46.40) and Ravens vs. Steelers (-54.50) suggest potential model issues, such as sensitivity to unscaled factors.
- **Centered Distributions**: Most games (e.g., Rams vs. Vikings at -0.94) are centered near 0, aligning with win probabilities near 0.5, indicating high uncertainty or balanced matchups.

3.3 Comparison with Actual Outcomes

The predicted win/loss outcomes were compared to the actual results of the 2025 NFL Wild Card games:

- Buffalo Bills vs. Denver Broncos: Predicted Win (Actual: Win, Correct).

- Baltimore Ravens vs. Pittsburgh Steelers: Predicted Loss (Actual: Loss, Correct).
- Houston Texans vs. Los Angeles Chargers: Predicted Win (Actual: Loss, Incorrect).
- Philadelphia Eagles vs. Green Bay Packers: Predicted Win (Actual: Loss, Incorrect).
- Tampa Bay Buccaneers vs. Washington Commanders: Predicted Loss (Actual: Loss, Correct).
- Los Angeles Rams vs. Minnesota Vikings: Predicted Loss (Actual: Loss, Correct).

The model correctly predicted 4 out of 6 games, achieving an accuracy of 66.67%. This indicates moderate predictive power, with errors likely due to sensitivity to unscaled factors and missing defensive metrics.

4 Discussion

4.1 Analysis of Results

The model predicts a mix of home wins and losses, with mean point differentials ranging from -54.50 to 46.40. Extreme values (e.g., Bills vs. Broncos, Ravens vs. Steelers) suggest sensitivity to unscaled factors, as seen in the wide violin plot distributions. Most games have probabilities near 0.5 and centered distributions, indicating balanced matchups or model limitations. The 66.67% accuracy against actual outcomes shows potential, but errors in games like Texans vs. Chargers and Eagles vs. Packers suggest areas for improvement.

4.2 Challenges

- **Team Mismatches**: Initial `NA` entries due to inconsistent abbreviations (e.g., `buffalo` vs. `buf`) were fixed by normalizing team names.
- **Factor Selection**: Removing `home/away-win-streak-last-3` simplified the model but may have omitted trends. Using only offensive and performance metrics might miss defensive impacts, contributing to extreme predictions.
- **Unscaled Factors**: The wide range of factor scales (e.g., `spread-line` from 0 to 15, `off-epa-per-play` from 0 to 0.5, `cpoe` from 0 to 1, `wpa` from -0.5 to 0.5, `home_points_last_3` and `away_points_last_3` from 0 to 150) likely amplifies extreme predictions.

5 Conclusion

This project developed a hierarchical Bayesian MCMC model to predict 2025 NFL Wild Card game outcomes, achieving a 66.67% accuracy in win/loss predictions. However, extreme predictions highlight the need for better factor scaling and additional metrics.

Future work includes:

- Normalizing factors (e.g., `off-epa-per-play`, `cpoe`, `wpa`, `home_points_last_3`, `away_points_last_3`) to reduce extreme predictions.
- Including defensive stats (e.g., `def-epa-per-play`) for a more complete model.
- Adding credible intervals to quantify uncertainty.
- Validating predictions against actual 2025 outcomes.
- Testing different prior variances for β_i and factor coefficients.

References

- NFL Data Source: `NFL-DataScrap(2024-2025).csv`, 2025.
- Hierarchical Bayesian Modeling with MCMC by Kenny Shirley - Data Mining, Columbia University November 17, 2011
- Gelman, A., et al. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.