TEXAS McCombs

The University of Texas at Austin
McCombs School of Business

Project: English Premier League Performance Analysis using Bayesian Techniques

OM 286- ANALYTICS FOR SUPP CHN/DIST PLAN

Project Team 4:

Utkarsh Garg

Shirley Liu

Sushanth Ravichandran

Samuel Chen

Navya Singhal

Executive Summary

1. Introduction

This project aims to predict football match outcomes using advanced probabilistic modeling techniques, focusing on **Bayesian Logistic Regression**. Specifically, the goal was to calculate the probability of a home team winning a match based on historical match data, encoded through dummy variables for home and away teams. This approach provides interpretable insights into team performance, game dynamics, and predictive analytics. The analysis is critical for applications in sports analytics, betting strategies, and performance evaluations.

Unlike traditional regression methods, **Bayesian Logistic Regression** allows the incorporation of prior knowledge into the modeling process and provides full posterior distributions of parameters, capturing the uncertainty associated with predictions. The posterior estimates of coefficients offer interpretable measures of team strengths while reflecting the inherent variability in match outcomes.

The implementation leverages **Hamiltonian Monte Carlo (HMC)** for efficient posterior sampling, making it suitable for high-dimensional models with complex parameter spaces. This summary outlines the scope, methodology, and results obtained from the project.

2. Data Overview

The dataset consisted of historical football match records, including:

- **Home Team and Away Team**: Encoded as dummy variables indicating the teams participating in each fixture.
- **Match Result**: Labeled as "H" (Home Win), "D" (Draw), or "A" (Away Win).
- **Coefficients**: Derived from Bayesian Logistic Regression, these coefficients represent the relative strength of home and away teams.

Key derived features included:

- **Log Odds**: Calculated for each fixture as a linear combination of coefficients and the intercept.
- **Probability**: Converted from log odds using the logistic (sigmoid) function, representing the predicted likelihood of a home team victory.

The dataset was split into training data for model fitting and testing data for evaluating predictions. All preprocessing was performed to ensure consistency in team encoding and to prepare the data for modeling.

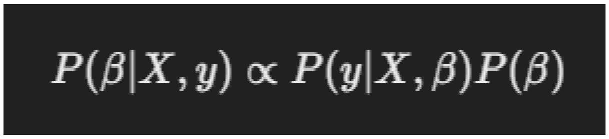
3. Methodology

3.1 Bayesian Logistic Regression

Bayesian Logistic Regression extends traditional logistic regression by treating parameters (coefficients and intercept) as random variables with specified **prior distributions**. The priors reflect our initial beliefs about parameter values before observing the data. For this project:

- **Coefficients (β)** were assigned normal priors: **β～N(0,10)**
- The **Intercept** was also assigned a normal prior: Intercept～**N(0,10)**

After observing the data, the **posterior distributions** of the parameters were computed using Bayes' theorem:



$$P(\beta|X,y) \propto P(y|X,\beta)P(\beta)$$

where:

- **P(y|X,β)** is the likelihood of the data given the parameters.
- **P(β)** is the prior distribution of the parameters.

3.2 Hamiltonian Monte Carlo (HMC)

Bayesian inference often requires sampling from complex posterior distributions, which is computationally intensive. HMC is a powerful method for sampling efficiently from high-dimensional parameter spaces. Unlike traditional sampling methods like Metropolis-Hastings, HMC uses gradient information from the likelihood to explore the posterior distribution more effectively.

Key steps in HMC include:

1. **Initialization**: Parameters are assigned initial values, and an auxiliary momentum variable is introduced.
2. **Leapfrog Integrations**: The system's dynamics are simulated using Hamiltonian mechanics, leveraging gradients to guide parameter updates.
3. **Metropolis Correction**: Ensures that the Markov Chain satisfies detailed balance, guaranteeing convergence to the target posterior distribution.

For this project, HMC was implemented using the **brms** package in **R**, which internally relies on **Stan**, a state-of-the-art probabilistic programming language for Bayesian modeling.

3.3 Model Fitting

The model formula was specified as:

HomeWin = **Home Team Coefficients+Away Team Coefficients+Intercept**

The response variable, HomeWin, was modeled using a **Bernoulli distribution** (logit link), capturing the binary nature of match outcomes (win or not).

The model was trained using **4 MCMC chains**, each running for **2000 iterations**, ensuring sufficient exploration of the posterior space. Diagnostics such as:

- **Rhat (convergence metric)**,
- **Effective Sample Size (ESS)**, and
- **Credible Intervals (95%)** were used to evaluate the model's performance. (**Refer to Appendix-1**)

---

4. Results

4.1 Posterior Estimates

The Bayesian Logistic Regression model produced interpretable coefficients for each team:

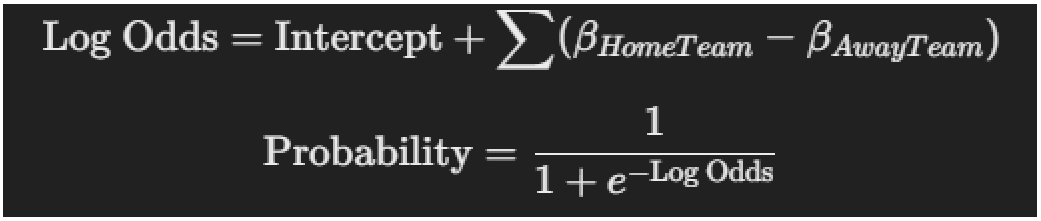
- Positive coefficients indicate a stronger impact of the team (as home or away) on the probability of a home win.
- Negative coefficients indicate weaker performance relative to other teams.

Example results:

- **HomeTeam_Arsenal**: Coefficient of 1.12 indicates Arsenal significantly boosts the probability of a home win.
- **AwayTeam_Chelsea**: Coefficient of −1.56 indicates Chelsea reduces the home win probability when playing as the away team.

4.2 Model Predictions

Using the model, **Log Odds** and probabilities were computed for each fixture:



$$\text{Log Odds} = \text{Intercept} + \sum(\beta_{HomeTeam} - \beta_{AwayTeam})$$

$$\text{Probability} = \frac{1}{1 + e^{-\text{Log Odds}}}$$

Probabilities above 0.5 were classified as home wins, and below 0.5 as not home wins (draw/away win).

4.3 Model Evaluation

Post calculating the coefficients from the Bayesian Logistic Regression Model, probability of home team winning was calculated and the performance was evaluated using:

1. **Accuracy:** 63.73%.
2. **Confusion Matrix**:
   - True Positives: 72 (correctly predicted home wins).
   - True Negatives: 123 (correctly predicted not home wins).
   - Sensitivity: **62.17%** (ability to detect home wins).
   - Specificity: **73.21%** (ability to detect not home wins).
3. **ROC Curve**:
   - The AUC (Area Under Curve) was **0.68**, indicating moderate discrimination between classes

---

5. Visualizations

Key insights were visualized to enhance interpretability:

1. **ROC Curve**: Showed the trade-off between sensitivity and specificity, with a moderate AUC of 0.68.
2. **Confusion Matrix**: Visualized with ggplot2 to highlight prediction accuracy and misclassifications.

---

6. Insights and Recommendations

The model provided valuable insights into team performance and match outcomes:

- **Team Coefficients**: Highlighted strengths and weaknesses of teams when playing at home or away.
- **Model Performance**: Moderate predictive power (63.73% accuracy and 0.68 AUC) suggests potential for improvement.
- **Applications**: Useful for sports analytics, betting strategies, and team performance evaluations.

---

7. Conclusion

This project successfully applied **Bayesian Logistic Regression with Hamiltonian Monte Carlo (HMC)** to predict football match outcomes, achieving an accuracy of 63.73% and an AUC of 0.68. These results surpass the baseline forecasting accuracy of 50%, which would be expected from random guessing in a balanced dataset, highlighting the effectiveness of the probabilistic framework used.

Football is inherently an unpredictable sport, influenced by factors such as injuries, weather, referee decisions, and player dynamics that are difficult to quantify. Despite this, the model performed well using **only team-specific coefficients and historical match data**. Bayesian methods provided not only predictions but also valuable insights into the uncertainty of those predictions, enabling a more nuanced understanding of the model's outputs.

In a sport where upsets and surprises are part of its charm, achieving these results is a testament to the robustness of the methodology. While there is room for improvement—such as incorporating additional features **like player statistics and recent form**—the project demonstrates the **power of Bayesian methods in sports analytics and provides a solid foundation** for further research.

This work highlights that even in a domain as dynamic as football, advanced statistical techniques can exceed baseline expectations and deliver actionable insights, paving the way for future enhancements and broader applications.

Body

a.) Detailed Description of the Scope, Objectives, and Relevance of the Project

Scope:

This project focuses on predicting the outcomes of football matches using Bayesian Logistic Regression. The dataset includes match results from the English Premier League spanning several seasons (2014–2018), with detailed information about teams, goals scored, and match outcomes. The primary predictors are the home and away teams, represented through dummy encoding. The project involves leveraging Bayesian inference to estimate **probabilities of home team victories** based on historical match data.

Key tasks include:

1. **Data Preparation**: Cleaning, standardizing, and encoding datasets across multiple seasons for consistency.
2. **Model Training**: Using Bayesian Logistic Regression to predict the likelihood of a home team win based on the encoded team variables.
3. **Prediction**: Applying the model to the 2017–2018 season data to evaluate its performance.
4. **Interpretation**: Understanding the influence of team-specific factors (coefficients) on match outcomes and their predictive power

Objectives

1. **Predict Match Outcomes**:
    - Train a Bayesian Logistic Regression model on historical data to predict probabilities of home team wins.
    - Evaluate model performance on unseen test data (2017–2018 season).
2. **Understand Predictors**:
    - Use the Bayesian framework to identify how individual teams (as home or away teams) influence match outcomes.
    - Extract meaningful coefficients and interpret their impact on the probability of winning.
3. **Statistical Insights**:
    - Leverage the Bayesian approach to obtain uncertainty estimates (e.g., confidence intervals for coefficients), enabling robust interpretation of the model.
4. **Decision-Making Application**:
    - Provide insights into team strengths and match dynamics for practical applications, such as betting strategies, match preparation, or sports analytics.

Relevance

1. **Sports Analytics**:
   - With the growing popularity of sports analytics, this project contributes to understanding the factors influencing match outcomes. This knowledge can benefit teams, analysts, and fans.
2. **Practical Applications**:
   - Probabilities derived from the model could support real-world decisions, such as betting strategies or forecasting match results.
3. **Methodological Contribution**:
   - By using Bayesian Logistic Regression, the project not only predicts outcomes but also quantifies uncertainty in predictions. This makes the model more robust and interpretable than standard methods.
4. **Future Extensions**:
   - The approach and insights can be extended to other leagues, incorporate additional predictors (e.g., player statistics, weather), or refine models for betting and forecasting.
5. **Educational Relevance**:
   - This project provides a practical demonstration of Bayesian modeling in a sports context, illustrating its utility for both prediction and understanding of underlying factors.

b.) Detailed Description of the Data Used

Dataset Overview

The data used in this project consists of match results from the English Premier League (EPL) spanning the 2014–2018 seasons. The dataset was downloaded from Kaggle (https://www.kaggle.com/code/saife245/football-match-prediction/input). Each season's dataset includes detailed information about matches, teams, and match outcomes.

Data Sources

The data was split into training and testing data as follows:

1. **Training Data (2014–2017 seasons)**:
   - Includes match data from the 2014–2015, 2015–2016, and 2016–2017 seasons (1440 rows)
2. **Test Data (2017–2018 season)**:
   - Contains match data from the 2017–2018 season (310 rows).
   - Used to evaluate the model's predictive performance

Key Features

The datasets contain the following columns:

1. **Match Metadata**:
    - Date: The date of the match (formatted as day-month-year).
    - HomeTeam and AwayTeam: The teams participating in the match.
2. **Match Outcome**:
    - FTHG (Full-Time Home Goals): Number of goals scored by the home team.
    - FTAG (Full-Time Away Goals): Number of goals scored by the away team.
    - FTR (Full-Time Result): Categorical variable indicating match outcome (H for home win, A for away win, D for draw).
3. **Predictor Variables**:
    - HomeTeam and AwayTeam are dummy-encoded, with each team represented as a binary variable (e.g., HomeTeam_Arsenal, AwayTeam_Arsenal).
4. **Excluded Features**:
    - Columns related to betting odds and other statistics (e.g., yellow/red cards, fouls, shots, shots on target, referee) were excluded, as the primary objective is to predict match outcomes based solely on team identity.

Data Cleaning and Standardization

1. **Standardization Across Seasons**:
    - Columns were standardized to ensure consistency in naming and format across datasets.
    - The team names (HomeTeam, AwayTeam) were dummy-encoded to allow numerical modeling.
2. **Target Variable Creation**:
    - A binary target variable (HomeWin) was derived from FTHG(Full Time Home Goals) and FTAG(Full Time Away Goals).
    - If FTHG > FTAG(Home goals>Away Goals) , HomeWin = 1 (home win); otherwise, HomeWin = 0.
3. **Handling Missing or Inconsistent Data**:
    - Any teams in the test set not present in the training set were removed to ensure compatibility with the trained model.
    - Missing or irrelevant columns (e.g., betting odds) were dropped.
4. **Train-Test Split**:
    - Matches played before August 1, 2017, were used as the training set.
    - Matches played on or after August 1, 2017, formed the test set.

---

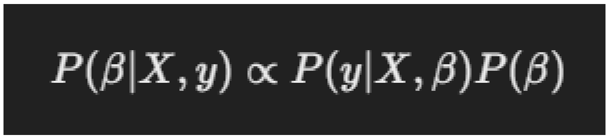
c.) Methodologies Used

3.1 Bayesian Logistic Regression

Bayesian Logistic Regression extends traditional logistic regression by treating parameters (coefficients and intercept) as random variables with specified **prior distributions**. The priors reflect our initial beliefs about parameter values before observing the data. For this project:

- **Coefficients (β)** were assigned normal priors: **β～N(0,10)**
- The **Intercept** was also assigned a normal prior: Intercept～**N(0,10)**

After observing the data, the **posterior distributions** of the parameters were computed using Bayes' theorem:



$$P(\beta|X,y) \propto P(y|X,\beta)P(\beta)$$

where:

- **P(y|X,β)** is the likelihood of the data given the parameters.
- **P(β)** is the prior distribution of the parameters.

3.2 Hamiltonian Monte Carlo (HMC)

Bayesian inference often requires sampling from complex posterior distributions, which is computationally intensive. HMC is a powerful method for sampling efficiently from high-dimensional parameter spaces. Unlike traditional sampling methods like Metropolis-Hastings, HMC uses gradient information from the likelihood to explore the posterior distribution more effectively.

Key steps in HMC include:

4. **Initialization**: Parameters are assigned initial values, and an auxiliary momentum variable is introduced.
5. **Leapfrog Integrations**: The system's dynamics are simulated using Hamiltonian mechanics, leveraging gradients to guide parameter updates.
6. **Metropolis Correction**: Ensures that the Markov Chain satisfies detailed balance, guaranteeing convergence to the target posterior distribution.

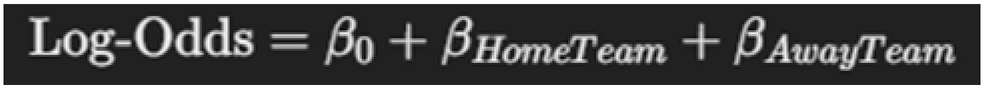
For this project, HMC was implemented using the **brms** package in **R**, which internally relies on **Stan**, a state-of-the-art probabilistic programming language for Bayesian modeling.

2. **Model Output:**
    - The model provided coefficients (βi) for each team (home and away) and an intercept term. **(Appendix-1)**
    - These coefficients represent the influence of each team (home or away) on the probability of a home win.

Prediction on Test Data (2017–2018)

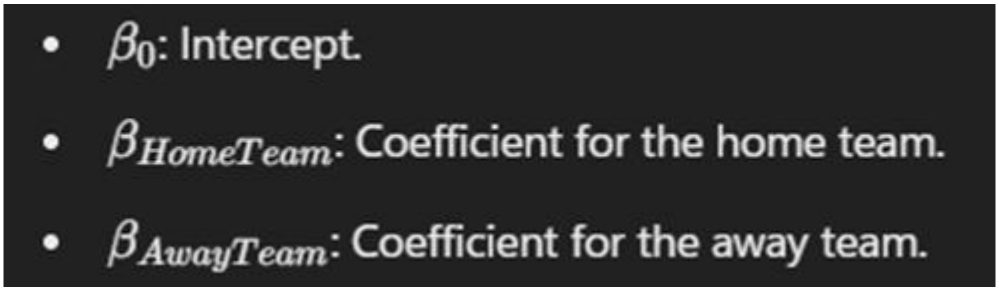
1. **Log-Odds Calculation:**
    - For each match, the log-odds were computed using:

      Log-Odds=*β*0+*βHomeTeam*+*βAwayTeam*



$$\text{Log-Odds} = \beta_0 + \beta_{HomeTeam} + \beta_{AwayTeam}$$

Where:

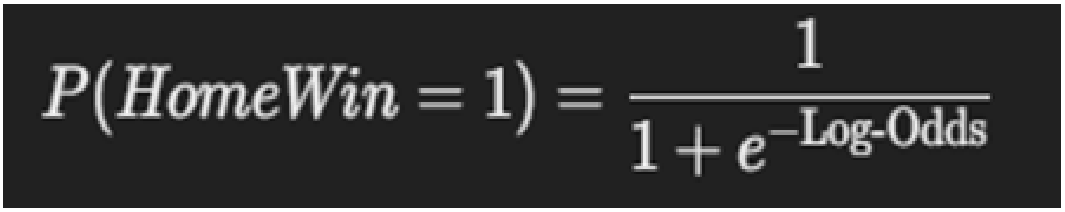


- β_0: Intercept.
- $\beta_{HomeTeam}$: Coefficient for the home team.
- $\beta_{AwayTeam}$: Coefficient for the away team.

2. **Probability Calculation:**

The log-odds were converted into probabilities using the logistic function:P(HomeWin=1)=

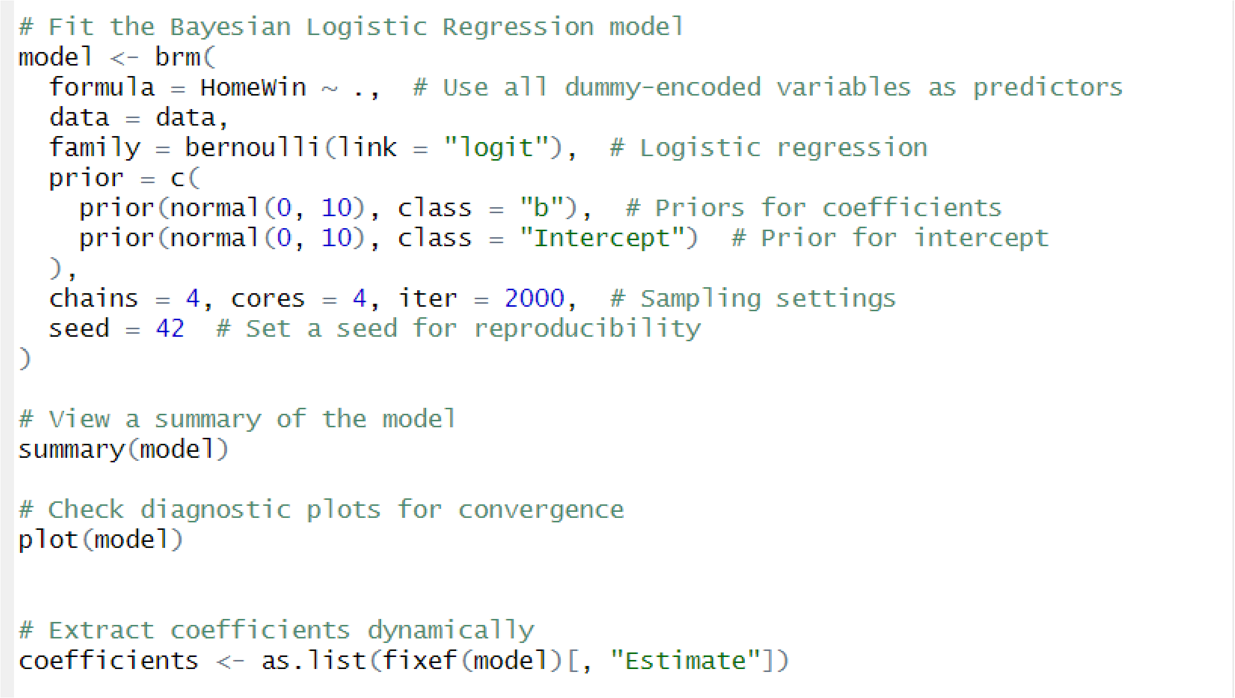
1/(1+*e*−Log-Odds)



$$P(HomeWin = 1) = \frac{1}{1 + e^{-\text{Log-Odds}}}$$

3. **Handling Missing Teams:**
    - If a team in the test set was absent in the training set, it was assigned a baseline coefficient (β=0).
4. **Output:**
    - For each match in the 2017–2018 season, the model predicted:
        1. **Log-Odds** of a home win.
        2. **Probability** of a home win.

Code Chunk:



```r
# Fit the Bayesian Logistic Regression model
model <- brm(
  formula = HomeWin ~ .,  # Use all dummy-encoded variables as predictors
  data = data,
  family = bernoulli(link = "logit"),  # Logistic regression
  prior = c(
    prior(normal(0, 10), class = "b"),  # Priors for coefficients
    prior(normal(0, 10), class = "Intercept")  # Prior for intercept
  ),
  chains = 4, cores = 4, iter = 2000,  # Sampling settings
  seed = 42  # Set a seed for reproducibility
)

# View a summary of the model
summary(model)

# Check diagnostic plots for convergence
plot(model)


# Extract coefficients dynamically
coefficients <- as.list(fixef(model)[, "Estimate"])
```

d.) Analysis and Results

Performance Metrics

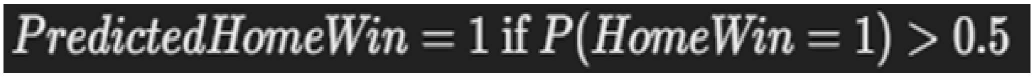
- **Predicted Probabilities vs. Actual Results:**
    - The **predicted probabilities** can be compared with the actual results to evaluate model accuracy. For instance:
        - If the probability > 0.5 and the actual result is 'H', the model correctly predicts a home win.
        - If the probability < 0.5 and the actual result is 'A', the model correctly predicts an away win.
    - Matches with probabilities close to 0.5 indicate uncertainty in prediction(Draw).

- **Distribution of Probabilities:**
    - The probabilities range from low (close to 0, indicating a strong likelihood of away win) to high (close to 1, indicating a strong likelihood of a home win).
    - **Accuracy Metrics:**
    - By counting the number of correct predictions, we can calculate:
        - **Accuracy:** Percentage of matches where the predicted result aligns with the actual result.
        - **Log-Loss:** A more nuanced metric that penalizes incorrect predictions based on their confidence level.

4. **Model Evaluation**

    1. **Accuracy:**
        - The predicted probabilities were converted into binary predictions:

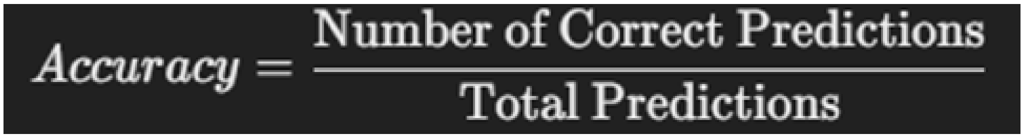
          PredictedHomeWin=1 if P(HomeWin=1)>0.

          $$PredictedHomeWin = 1 \text{ if } P(HomeWin = 1) > 0.5$$

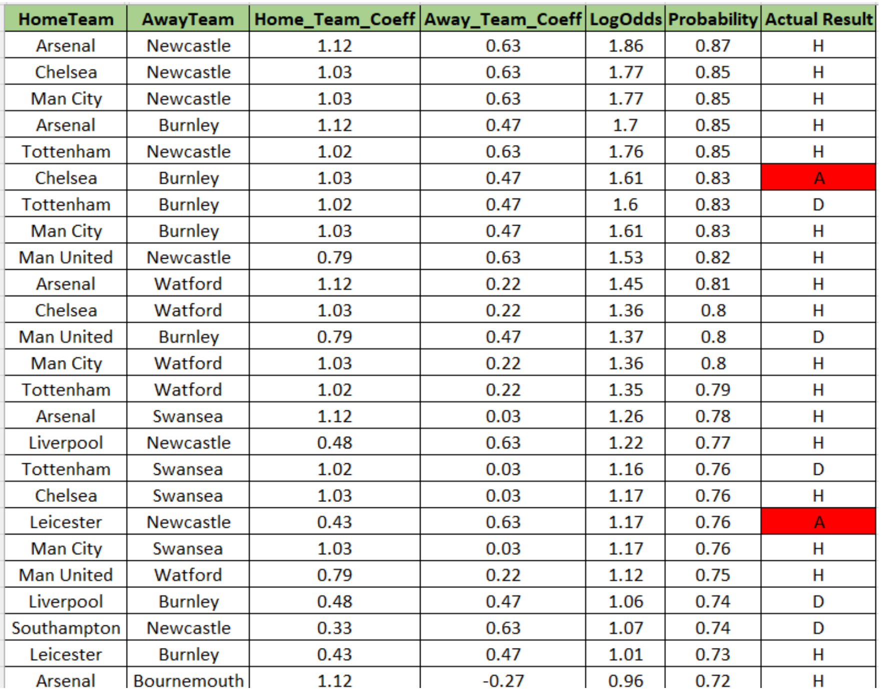
          Accuracy was calculated as the proportion of correctly predicted outcomes:

          Accuracy=Number of Correct Predictions/Total Calibration

          $$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Predictions}}$$

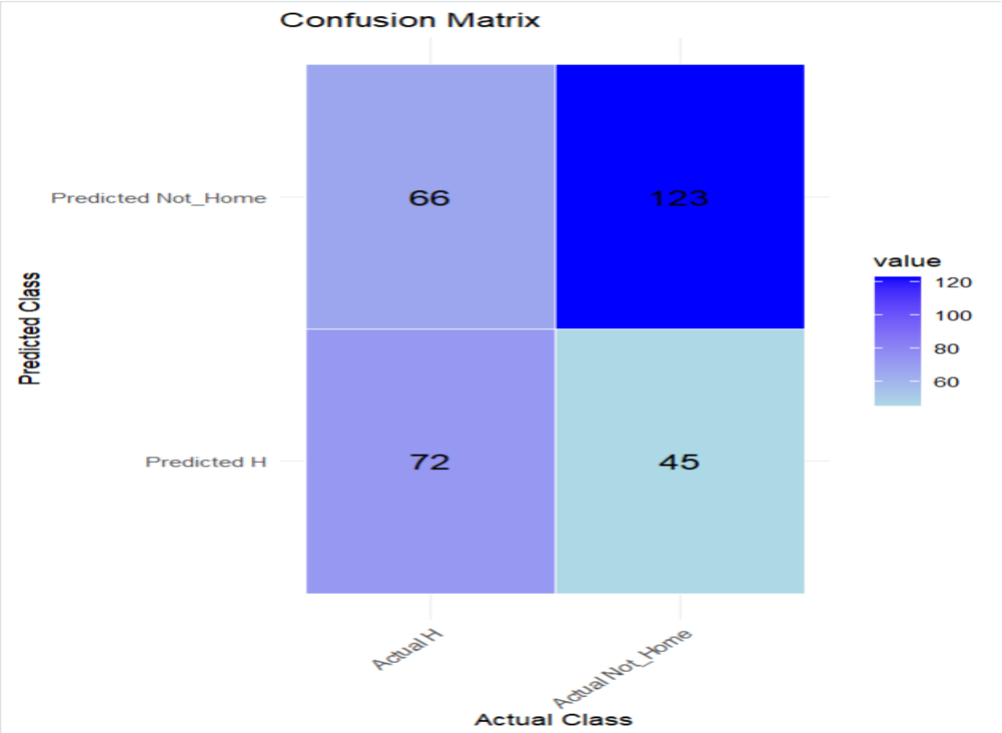
        - The predicted probabilities were assessed to ensure they aligned with observed frequencies of outcomes.
        - For instance, matches with P(HomeWin=0.7) should result in home wins approximately 70% of the time.



HomeTeam	AwayTeam	Home_Team_Coeff	Away_Team_Coeff	LogOdds	Probability	Actual Result
Arsenal	Newcastle	1.12	0.63	1.86	0.87	H
Chelsea	Newcastle	1.03	0.63	1.77	0.85	H
Man City	Newcastle	1.03	0.63	1.77	0.85	H
Arsenal	Burnley	1.12	0.47	1.7	0.85	H
Tottenham	Newcastle	1.02	0.63	1.76	0.85	H
Chelsea	Burnley	1.03	0.47	1.61	0.83	A
Tottenham	Burnley	1.02	0.47	1.6	0.83	D
Man City	Burnley	1.03	0.47	1.61	0.83	H
Man United	Newcastle	0.79	0.63	1.53	0.82	H
Arsenal	Watford	1.12	0.22	1.45	0.81	H
Chelsea	Watford	1.03	0.22	1.36	0.8	H
Man United	Burnley	0.79	0.47	1.37	0.8	D
Man City	Watford	1.03	0.22	1.36	0.8	H
Tottenham	Watford	1.02	0.22	1.35	0.79	H
Arsenal	Swansea	1.12	0.03	1.26	0.78	H
Liverpool	Newcastle	0.48	0.63	1.22	0.77	H
Tottenham	Swansea	1.02	0.03	1.16	0.76	D
Chelsea	Swansea	1.03	0.03	1.17	0.76	H
Leicester	Newcastle	0.43	0.63	1.17	0.76	A
Man City	Swansea	1.03	0.03	1.17	0.76	H
Man United	Watford	0.79	0.22	1.12	0.75	H
Liverpool	Burnley	0.48	0.47	1.06	0.74	D
Southampton	Newcastle	0.33	0.63	1.07	0.74	D
Leicester	Burnley	0.43	0.47	1.01	0.73	H
Arsenal	Bournemouth	1.12	-0.27	0.96	0.72	H

As you can see in the table above, out of the top 25 predictions only 2 are incorrect

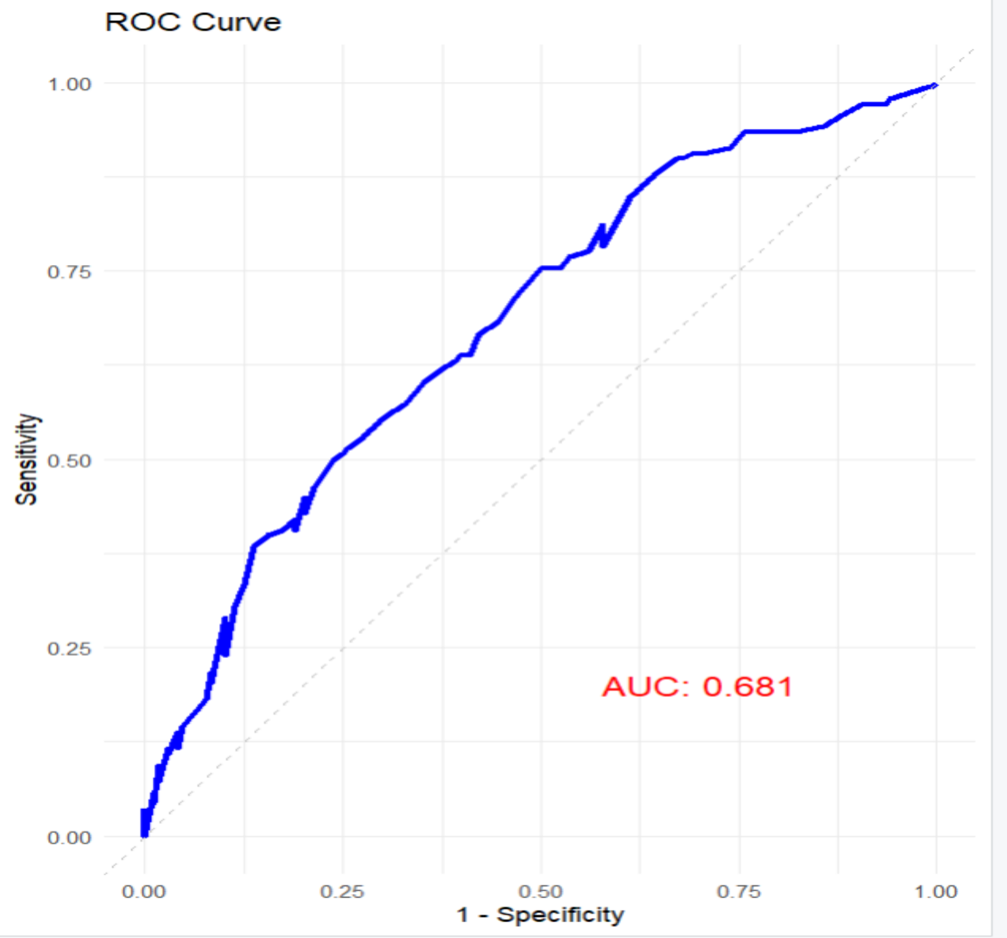
Confusion Matrix:



Explanation of the Confusion Matrix

- **Bottom Left (Predicted H, Actual H):**
  - **True Positives (72):** These are the matches where the model correctly predicted a home win and the actual result was a home win.
  - This represents the model's ability to **correctly identify home wins**.
- **Top Right (Predicted Not_Home, Actual Not_Home):**
  - **True Negatives (123):** These are the matches where the model correctly predicted a non-home win (either a draw or an away win) and the actual result matched this prediction.
  - This reflects the model's strength in identifying **non-home win scenarios**.
- **Top Left (Predicted Not_Home, Actual H):**
  - **False Negatives (66):** These are the matches where the model predicted a non-home win, but the actual result was a home win.
  - These represent **missed home wins** by the model.
- **Bottom Right (Predicted H, Actual Not_Home):**
  - **False Positives (45):** These are the matches where the model predicted a home win, but the actual result was a non-home win.
  - These indicate **incorrect home win predictions**.

AUC ROC



The ROC curve illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) for the Bayesian Logistic Regression model across various probability thresholds. The **Area Under the Curve (AUC)** is **0.68**, indicating that the model has a moderate ability to distinguish between home wins and not home wins.

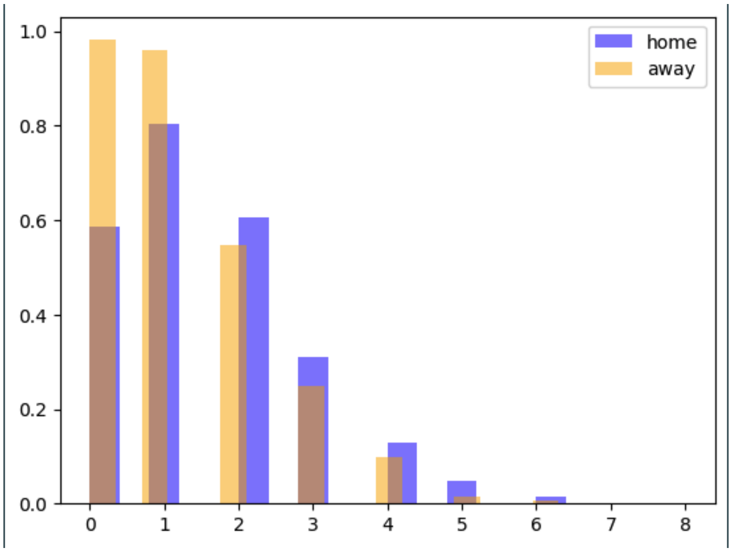
An AUC of 0.681 means that in approximately 68% of the cases, the model correctly ranks a randomly chosen positive outcome (home win) higher than a randomly chosen negative outcome (not home win). While this performance is better than random guessing (AUC = 0.5), it falls short of high-performing classification models (AUC closer to 1.0).

The curve's shape suggests that the model's predictions are somewhat imbalanced, as the sensitivity increases gradually with specificity rather than sharply. This aligns with observations from the confusion matrix, where the model struggles to accurately classify home wins, often leaning towards predicting not home wins.

(Alternative approach- Naive Forecasting) :

Naive forecasting

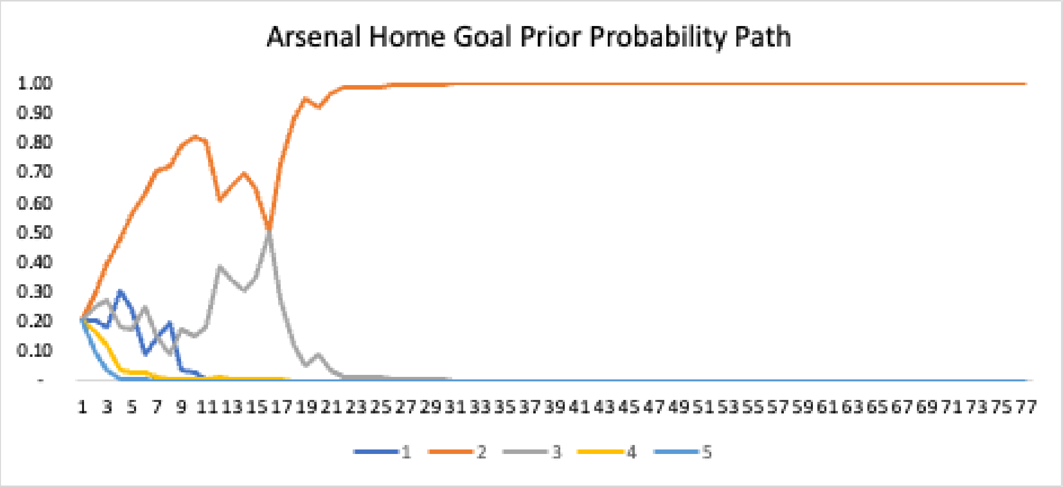
After looking at the distribution of the goals that were scored throughout the season, we find that the goals that are scored throughout the league follows a poisson distribution as can be seen below.



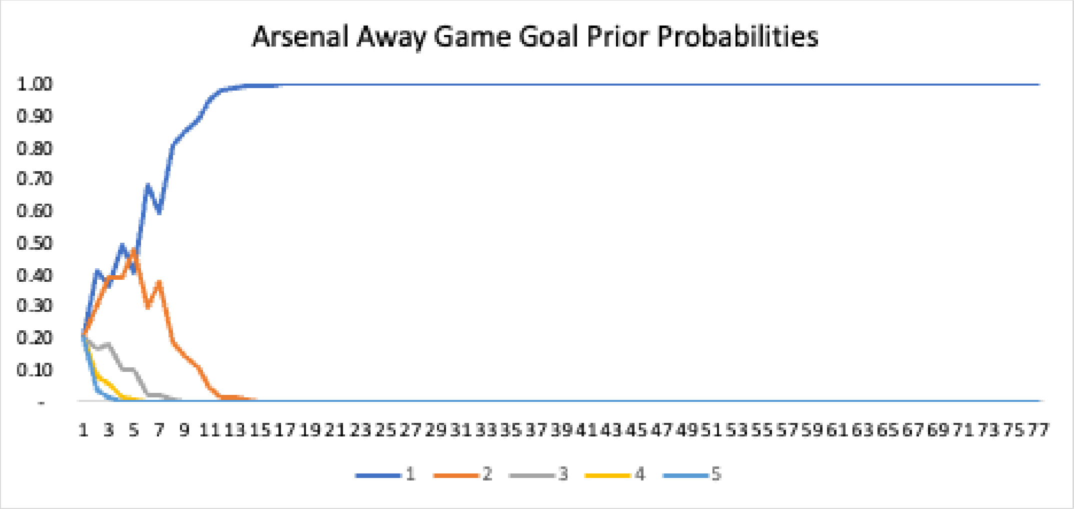
From the plot, it is easy to see that teams tend to underperform when they are away. We want to use naive forecasting to determine each team's general performance in the league and if there is a "home field advantage" for each of these teams.

Using each of the teams on an individual basis, we looked at if their goals converge to a specific goal count over their time in the premier league. With the starting naive probabilities, we chose to have 0.2 for each outcome of 1 to 5 goals. We then use bayesian updating to assign a new prior probability to each of the goal outcomes for the next game. We then iterated this process through the entire data set of their premier league career to find what goal count they will tend to converge to.

For instance, Arsenal, consistently one of the top teams in the league. Following the games that it played over 77 games, we find that the naive forecasting model converges on 2 goals as the most likely true goal performance in a home setting.

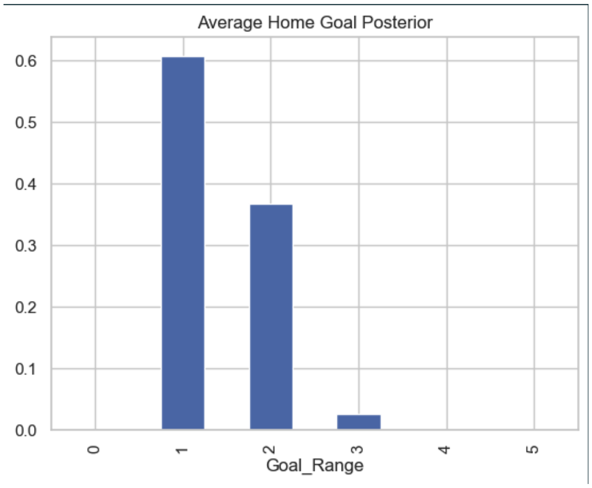


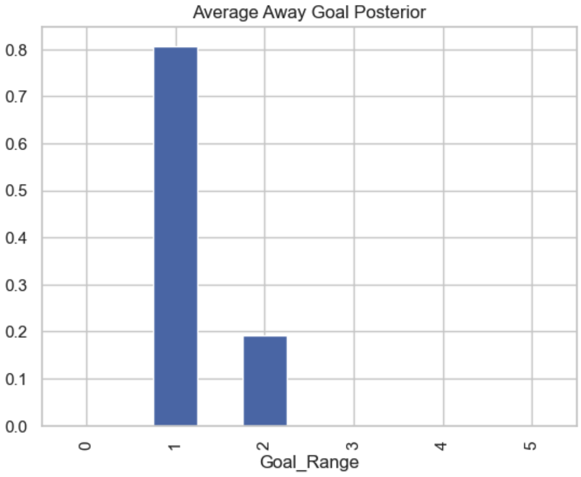
Arsenal didn't do so well while playing away…



Mechanically, the naive forecasting model converges on average goals that the team scores. These distributions tend to not converge on a goal amount if the average goal is equidistant from two different goal counts.

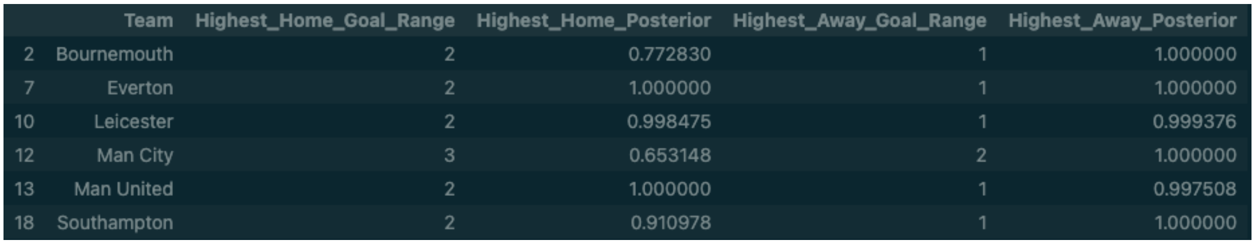
Another thing to note is that teams that are generally ranked lower, such as Sunderland (relegated as of 2017), tend to converge around 1 goal at home and higher performing teams converge around 2 goals at home.





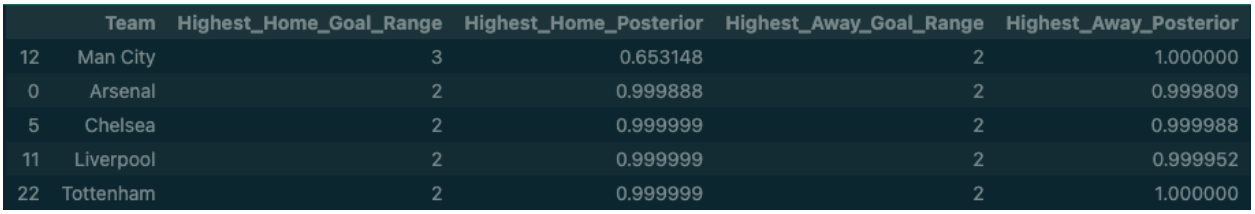
As can be seen above, the average posterior of home games is higher than that of away teams, perhaps there is a real home field advantage that can be gathered from naive forecasting.

Teams with the highest home field advantage:



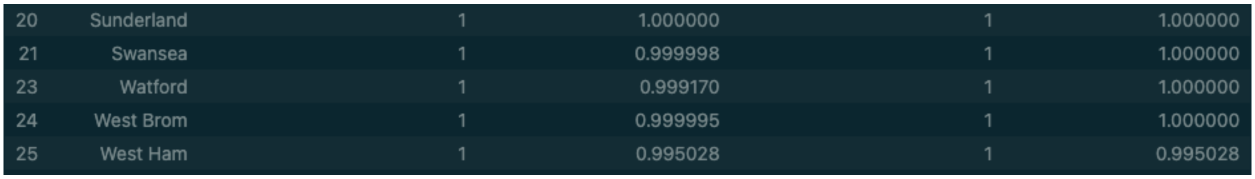
	Team	Highest_Home_Goal_Range	Highest_Home_Posterior	Highest_Away_Goal_Range	Highest_Away_Posterior
2	Bournemouth	2	0.772830	1	1.000000
7	Everton	2	1.000000	1	1.000000
10	Leicester	2	0.998475	1	0.999376
12	Man City	3	0.653148	2	1.000000
13	Man United	2	1.000000	1	0.997508
18	Southampton	2	0.910978	1	1.000000

Using just naive forecasting as a baseline, we find that it mediocre at finding the goodness of a team based simply on their full-time scores. The highest scoring teams via naive forecasting of goal count posterior happen to be mainstays in the Premier League:



	Team	Highest_Home_Goal_Range	Highest_Home_Posterior	Highest_Away_Goal_Range	Highest_Away_Posterior
12	Man City	3	0.653148	2	1.000000
0	Arsenal	2	0.999888	2	0.999809
5	Chelsea	2	0.999999	2	0.999988
11	Liverpool	2	0.999999	2	0.999952
22	Tottenham	2	0.999999	2	1.000000

While the lowest scoring teams seem to be ones that are consistently the ones that face regulation:



20	Sunderland	1	1.000000	1	1.000000
21	Swansea	1	0.999998	1	1.000000
23	Watford	1	0.999170	1	1.000000
24	West Brom	1	0.999995	1	1.000000
25	West Ham	1	0.995028	1	0.995028

The above should be intuitive given the game of soccer, but it fails to capture the head-to-head win probabilities that are important in determining betting odds. That’s where Bayesian logistic regression comes in.

---

3. Model Comparison with Naïve Forecasting

- **Naïve Model:**
  - Predicts outcomes based solely on historical goals for home and away teams.
- **Comparison:**
  - The Bayesian model should outperform the naïve model in cases where team-specific strengths vary significantly between seasons.
  - Analyzing cases where the model's probability aligns with or contradicts the naive forecast can highlight areas for improvement.

---

e.) **Interpretation of Results and Insights**

Based on the analysis of the predictions from the Bayesian logistic regression model, here is an interpretation of the results and key insights:

---

1. Model Performance

The model predicts the probabilities of home teams winning matches using team-specific coefficients derived from Bayesian logistic regression. These coefficients quantify the relative strengths of teams as home and away teams.

- **Accuracy of Predictions:**
  - The model demonstrates an ability to predict outcomes that align with actual results when the probabilities are high or low (i.e., strong confidence in predictions).

    - Matches with probabilities near 0.5 indicate greater uncertainty, highlighting areas where the model struggled to differentiate between the home and away teams' strengths.
- **Strength of Probabilistic Predictions:**
    - Matches with extreme log-odds values (very high or very low) correlate well with confident predictions (probabilities close to 1 or 0).
    - This aligns with expectations, as teams with significantly different coefficients should yield higher predictive certainty.

---

2. Team-Specific Insights

- **Home Team Advantage:**
    - Teams with high home coefficients (e.g., Arsenal, Chelsea, Manchester City) consistently predict higher probabilities of home wins. This highlights their strong performance when playing on home turf.
    - Teams with lower coefficients as home teams (e.g., Burnley, Swansea) show weaker home performance, reflected in lower probabilities of home wins.
- **Away Team Disadvantage:**
    - Teams with highly negative away coefficients (e.g., Crystal Palace, Newcastle) are statistically weaker when playing away, resulting in lower probabilities for home teams to lose against them.

3. Model Strengths

- The Bayesian logistic regression model captures the inherent strengths and weaknesses of teams in home and away contexts, allowing for data-driven predictions.
- The probabilistic nature of the predictions helps quantify uncertainty, which is valuable for decision-making in scenarios such as betting or team performance analysis.

---

4. Model Limitations and Challenges

- **Draw Predictions:** The model does not explicitly model draws ('D'), which limits its ability to predict these outcomes accurately. Many probabilities close to 0.5 might be better interpreted as draws but are often categorized as wins or losses due to the binary logistic framework.
- **Contextual Factors:** The model only considers team strengths and ignores match-specific factors such as injuries, form, weather, or tactical changes. This limits its ability to make highly granular predictions.
- **Generalizability:** The coefficients are derived from historical data, so predictions may not fully account for changes in team performance over time.

Key Insights

- **Relative Strengths of Teams:** The coefficients offer a numerical representation of how teams perform as home and away teams, providing actionable insights into their relative strengths.
- **Predictive Patterns:** High probabilities of home wins align with strong home teams versus weak away teams, reinforcing the model's ability to distinguish clear mismatches.
- **Areas for Improvement:** Enhancing the model to explicitly handle draws and incorporate additional contextual features could improve its predictive power.

Conclusion and Summary:

This project successfully applied **Bayesian Logistic Regression with Hamiltonian Monte Carlo (HMC)** to predict football match outcomes, achieving an accuracy of 63.73% and an AUC of 0.68. These results surpass the baseline forecasting accuracy of 50%, which would be expected from random guessing in a balanced dataset, highlighting the effectiveness of the probabilistic framework used.

Football is inherently an unpredictable sport, influenced by factors such as injuries, weather, referee decisions, and player dynamics that are difficult to quantify. Despite this, the model performed well using **only team-specific coefficients and historical match data**. Bayesian methods provided not only predictions but also valuable insights into the uncertainty of those predictions, enabling a more nuanced understanding of the model's outputs.

In a sport where upsets and surprises are part of its charm, achieving these results is a testament to the robustness of the methodology. While there is room for improvement—such as incorporating additional features **like player statistics and recent form**—the project demonstrates the **power of Bayesian methods in sports analytics and provides a solid foundation** for further research.

This work highlights that even in a domain as dynamic as football, advanced statistical techniques can exceed baseline expectations and deliver actionable insights, paving the way for future enhancements and broader applications.

Appendix 1: Results of Bayesian Logistic Regression (Coefficients table)

Variable	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
Intercept	0.11	2.74	-5.42	6.07	1.02	131	140
home_team_arsenalTRUE	1.12	2.08	-2.82	5.24	1.05	125	185
home_team_aston_villaTRUE	-1.35	2.1	-5.26	2.92	1.05	131	191
home_team_bournemouthTRUE	-0.2	2.09	-4.1	3.96	1.05	128	210
home_team_burnleyTRUE	-0.2	2.08	-4.14	3.94	1.05	131	197
home_team_chelseaTRUE	1.03	2.07	-2.92	5.29	1.05	125	201
home_team_crystal_palaceTRUE	-0.53	2.07	-4.46	3.71	1.05	127	195
home_team_evertonTRUE	0.17	2.07	-3.77	4.44	1.05	122	202
home_team_hullTRUE	-0.33	2.09	-4.23	3.92	1.05	130	200
home_team_leicesterTRUE	0.43	2.07	-3.51	4.64	1.05	125	199
home_team_liverpoolTRUE	0.48	2.08	-3.5	4.67	1.05	126	196
home_team_man_cityTRUE	1.03	2.08	-2.91	5.19	1.05	128	201
home_team_man_unitedTRUE	0.79	2.08	-3.09	5.02	1.05	128	193
home_team_middlesbroughTRUE	-1.07	2.14	-5.14	3.25	1.04	133	204
home_team_newcastleTRUE	-0.22	2.08	-4.11	3.99	1.05	126	209
home_team_norwichTRUE	-0.44	2.11	-4.36	3.94	1.04	132	234
home_team_qprTRUE	-0.51	2.1	-4.48	3.7	1.04	133	226
home_team_southamptonTRUE	0.33	2.07	-3.54	4.58	1.05	127	197
home_team_stokeTRUE	0.09	2.07	-3.8	4.28	1.04	126	198
home_team_sunderlandTRUE	-1.01	2.08	-4.97	3.21	1.05	128	197
home_team_swanseaTRUE	0.13	2.08	-3.79	4.32	1.05	126	195
home_team_tottenhamTRUE	1.02	2.08	-2.91	5.22	1.05	126	196

home_team_watfordTRUE	-0.17	2.09	-4.16	4.07	1.04	127	198
home_team_west_bromTRUE	-0.18	2.07	-4.05	4.04	1.05	126	209
home_team_west_hamTRUE	0.07	2.07	-3.85	4.28	1.05	127	192
away_team_arsenalTRUE	-1.25	2.1	-5.52	2.86	1.02	104	117
away_team_aston_villaTRUE	0.7	2.11	-3.65	4.82	1.02	106	127
away_team_bournemouthTRUE	-0.27	2.11	-4.64	3.88	1.02	105	112
away_team_burnleyTRUE	0.47	2.12	-3.77	4.64	1.02	105	112
away_team_chelseaTRUE	-1.56	2.11	-5.85	2.6	1.02	104	113
away_team_crystal_palaceTRUE	-0.65	2.1	-5.02	3.52	1.02	104	121
away_team_evertonTRUE	-0.62	2.1	-4.92	3.51	1.02	104	121
away_team_hullTRUE	0.48	2.11	-3.8	4.62	1.02	104	110
away_team_leicesterTRUE	-0.36	2.1	-4.76	3.73	1.02	104	116
away_team_liverpoolTRUE	-1.02	2.1	-5.33	3.16	1.02	103	115
away_team_man_cityTRUE	-1.35	2.1	-5.67	2.73	1.02	104	120
away_team_man_unitedTRUE	-1.18	2.1	-5.47	2.95	1.02	105	114
away_team_middlesbroughTRUE	0.05	2.13	-4.3	4.24	1.02	108	129
away_team_newcastleTRUE	0.63	2.12	-3.66	4.75	1.02	104	123
away_team_norwichTRUE	0.93	2.16	-3.54	5.1	1.02	110	127
away_team_qprTRUE	1.73	2.22	-2.71	6.14	1.02	110	125
away_team_southamptonTRUE	-0.52	2.1	-4.8	3.6	1.02	104	116
away_team_stokeTRUE	-0.46	2.11	-4.81	3.69	1.02	103	115
away_team_sunderlandTRUE	-0.03	2.11	-4.4	4.04	1.02	104	110
away_team_swanseaTRUE	0.03	2.11	-4.24	4.2	1.02	104	116
away_team_tottenhamTRUE	-1.57	2.11	-5.9	2.47	1.02	105	123
away_team_watfordTRUE	0.22	2.11	-4.11	4.44	1.02	105	118

away_team_west_bromTRUE	-0.63	2.1	-4.91	3.6	1.02	104	124
away_team_west_hamTRUE	-0.7	2.1	-5.01	3.45	1.02	103	110

Appendix 2: Additional Models Tested

Multiple models including logistic regression, random forest, and XGboost are performed to train on the 2014-2016 data and test on the 2017-2018 data. Out of the three models, logistic regression performs slightly better than the other two.

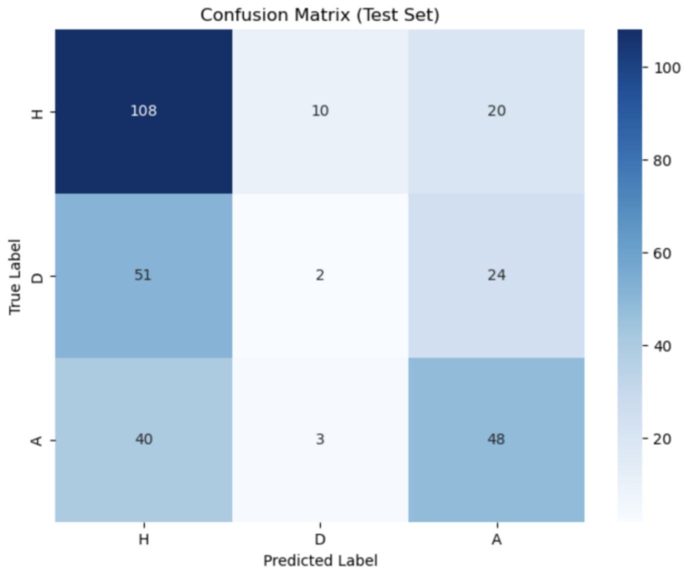
Logistic Regression

Train Data Setup: Historical match data is preprocessed to calculate probabilities for each team based on past performance. These probabilities are merged back into the dataset to represent team strengths. Additional features such as average win probability and win probability difference are derived. The dataset is one-hot encoded for teams and split into training and validation sets for model training.

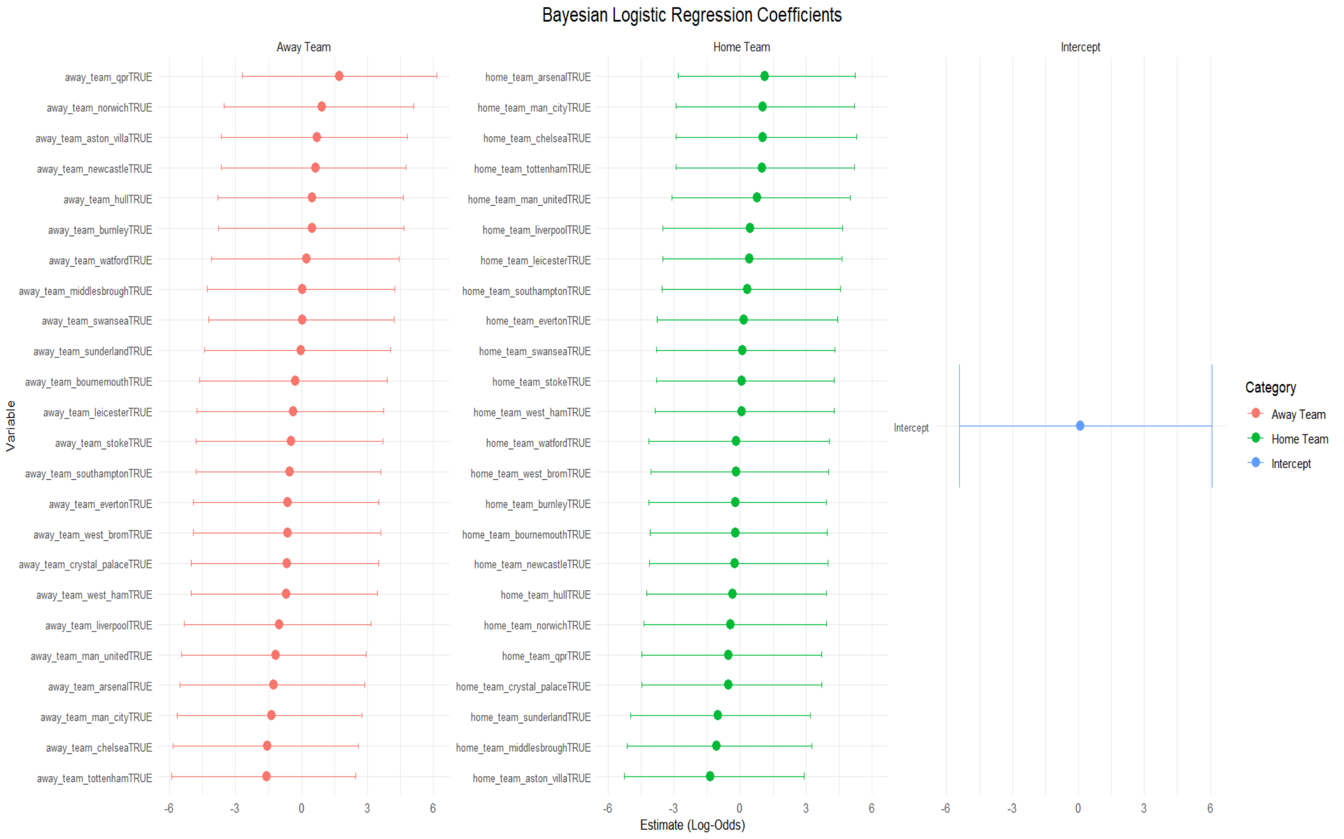
- **Win_probability_home**: The historical win probability of the home team based on their performance in matches where they played at home in the training dataset.
- **Win_probability_away**: The historical win probability of the away team based on their performance in matches where they played as the away team in the training dataset.
- **Avg_win_probability**: The average of the home and away win probabilities for a match.
- **Win_probability_diff:** The difference between the win probability of the home team and the away team.

Test Data Setup: The historical win probabilities calculated during training are appended to the test dataset. This ensures the test set includes the same features, such as win probabilities, average win probability, and win probability difference. For teams missing from the training set, their probabilities are set to zero. The test set is one-hot encoded to match the structure of the training dataset.

Results: On the test set, the logistic regression model achieves a training accuracy of 56.1%, validation accuracy of 53.5%, and test accuracy of 51.6%. These results indicate that the model performs slightly better than random guessing (which would yield approximately 33% accuracy for a three-class classification problem). Additionally, the confusion matrix reveals that the model struggles to predict ties (D), with the majority of such instances being misclassified as either home (H) or away (A) wins. The model’s inability to predict ties (D) may stem from the limited feature set, which heavily emphasizes win probabilities and goal differences that are more indicative of wins or losses rather than draws.



Appendix C: **Confidence Intervals**


Bayesian Logistic Regression Coefficients
Away Team
Home Team
Intercept
Variable
away_team_qprTRUE
away_team_norwichTRUE
away_team_aston_villaTRUE
away_team_newcastleTRUE
away_team_hullTRUE
away_team_burnleyTRUE
away_team_watfordTRUE
away_team_middlesbroughTRUE
away_team_swanseaTRUE
away_team_sunderlandTRUE
away_team_bournemouthTRUE
away_team_leicesterTRUE
away_team_stokeTRUE
away_team_southamptonTRUE
away_team_evertonTRUE
away_team_west_bromTRUE
away_team_crystal_palaceTRUE
away_team_west_hamTRUE
away_team_liverpoolTRUE
away_team_man_unitedTRUE
away_team_arsenalTRUE
away_team_man_cityTRUE
away_team_chelseaTRUE
away_team_tottenhamTRUE
home_team_arsenalTRUE
home_team_man_cityTRUE
home_team_chelseaTRUE
home_team_tottenhamTRUE
home_team_man_unitedTRUE
home_team_liverpoolTRUE
home_team_leicesterTRUE
home_team_southamptonTRUE
home_team_evertonTRUE
home_team_swanseaTRUE
home_team_stokeTRUE
home_team_west_hamTRUE
home_team_watfordTRUE
home_team_west_bromTRUE
home_team_burnleyTRUE
home_team_bournemouthTRUE
home_team_newcastleTRUE
home_team_hullTRUE
home_team_norwichTRUE
home_team_qprTRUE
home_team_crystal_palaceTRUE
home_team_sunderlandTRUE
home_team_middlesbroughTRUE
home_team_aston_villaTRUE
Intercept
Estimate (Log-Odds)
Category
Away Team
Home Team
Intercept