ISYE 6420 - Project 1

Michael J. Paglia gtID: 904052187 November 28, 2024

Introduction

0.1 Motivation

The prediction of football (soccer) matches presents unique challenges due to the low scoring nature of the sport and the interplay of various performance metrics, such as expected goals (xG), chances created, possession, interceptions, tackles, and more. Traditional models, such as solely relying on a Poisson distribution, fail to capture the temporal dynamics of goal scoring throughout the 90-minute period and the relative importance, or weight, of different statistical indicators. My model successfully incorporates various advanced metrics with Weibull distributions and Bayesian inference to arrive at match predictions for any of the 20 teams in the $2024-25$ English Premier League (EPL).

0.2 Features

Goals are not uniformly distributed across match time, with the average first goal occurring around the 29th minute and subsequent goals following a non-linear pattern[1]. Evidently, these patterns would be especially different for perceived stronger and weaker teams. These patterns are extracted via:

- Weibull-hazard function for time-dependent goal scoring[2]
- Derived "team strength" parameters based on offensive (e.g., percentile of goals scored) and defensive metrics (e.g. percentile of goals against)
- Normalized priors based on league averages

0.3 APIs and Libraries Used

- `soccerdata (FBRef)` [3] API used to fetch statistics for team and player metrics

[1] "In 120 matches played in total this season, there have been 297 goals scored between the 16th and the 90th minute, averaging 2.48 per match. In the 80 matches with no goal scored in the first 15 minutes, there have been 210 goals scored between the 16th and the 90th minute, averaging 2.63 per match. In the 40 matches with at least one goal scored in the first 15 minutes, there have been 87 goals scored between the 16th and the 90th minute, averaging 2.18 per match." from https://www.soccerstats.com/stats.asp?page=10

[2] A Weibull renewal process is used to model "time until a goal is scored" while accounting for the renewal, e.g., accumulation of a goal opportunity over the course of 90 minutes with a reset once a goal is scored. from https://medium.com/analytics-vidhya/distribution-of-premier-league-goals-855c909c6955

[3] from https://soccerdata.readthedocs.io/en/latest/datasources/FBref.html

- `Football Web Pages`[4] used to fetch statistics for league table statistics
- `pandas` data manipulation and analysis, particularly in handling team statistics and match data as DataFrames
- `numpy` numerical computing capabilities, used for mathematical operations and array manipulations throughout the model
- `arviz` statistical analysis and visualization; analyzing and interpreting the PyMC sampling results
- `Matplotlib` and `Seaborn` used to create visualizations of model results and statistical distributions
- `PyMc` implements the Bayesian model, including the Weibull hazard function and MCMC sampling

0.4 Relevant Files

- `API_KEY.txt` authentication key 4; anyone is free to use this, as it incurs no cost
- `epl_teams.xlsx` Excel file with Premier League team rosters and player information as per game week (GW) 1, scraped separately3
- `config.py` configuration file storing constants, statistical categories, and team name mappings used throughout the project
- `football_web_pages_endpoint.py` helper method to fetch current (Gw11) league table data and team statistics from endpoint
- `prepare_match.py` "main" file with CLI and Bayesian statistical analysis using PyMC sampling
- `prepare_rosters.py` helper methods to process the Excel roster file and prepares team-player mappings for use in the game simulation
- `trace_analysis.py` various functions for analyzing PyMC sampling results, visualizes posterior distributions and calculation of match probabilities, including weights of metrics used in simulation

[4] from https://www.footballwebpages.co.uk/api

Preliminaries

0.5 Performance Metrics

The model outlined in this project makes use of the following offensive and defensive metrics to derive a tensor representative of a team's "strength".

0.5.1 Offensive Metrics (Definitions)

- `Goals scored`
- `Expected goals (xG) per 90 minutes` measures the quality of a chance by calculating the likelihood that it will be scored, using information on similar shots in the past [5]
- `Expected assisted goals (xAG) per 90 minutes` measures a player's ability to set up scoring chances without having to rely on the actual result of the shot or the shooter's luck/ability [6]

0.5.2 Defensive Metrics (Definitions)

- `Tackles won` are when the tackler or one of their teammates regains possession because of the challenge, or when the ball goes out of play and is safe 5
- `Interceptions` are when a player reads an opponent's pass and intercepts the ball by moving into the line of the intended pass 5
- `Shots blocked` is an attempt to score by either a) attempt on target that is blocked by an outfield player, where other defenders or a goalkeeper are behind the 'blocker' or b) a shot blocked unintentionally by the shooter's own teammate 5
- `Clearances` are last-line shots blocked 5
- `Challenge success rate` involves either a tackle or use of body contact to regain possession successfully for the team

The parameters outlined above were normalized for each team, relative to all other teams in the league. Possession for each team is also counted, but cannot be considered either an offensive or defensive metric in particular. An extra step was taken to calculate how much each player on a team is responsible for these metrics to extract their impact on success. For instance, how much does Mohamed Salah of Liverpool contribute to his team's success? How are Liverpool affected if he is not in the starting lineup?

[5] from https://theanalyst.com/2024/07/opta-football-stats-definitions
[6] from https://fbref.com/en/expected-goals-model-explained/

Methodology

The model utilizes Bayesian inference and the empirical data referenced previously. The main components are as follows.

0.6 Data Processing and Normalization

Team and player statistics are retrieved from FBref using the `soccer data` Python library. They are normalized against league maximums for comparable scales, for instance, properly reflecting a team rated first in goals scored versus the team last in goals scored. This includes offensive metrics (goals, xG, xAG) and defensive metrics (tackles, interceptions, blocks, clearances, challenge success). Each statistic is processed by:

$$\frac{team\_statistic}{league\_maximum\_statistic}$$

0.7 Team Strength

Team strengths are calculated by aggregating individual player statistics and incorporating team-level metrics. The model uses two primary strength components:

- **Attack Strength**: Weighted aggregation of normalized goals, expected goals (xG), and expected assisted goals (xAG)
- **Defense Strength**: Weighted aggregation of tackles won, interceptions, shots blocked, clearances, and challenge success rates

Weighted combinations are used w.r.t. attack and defense strength in an effort to determine team performance more comprehensively. The weights for each metric are learned by informative priors centered on league averages with a fairly low value for σ (less than 1) to allow for some degree of variance. For numerical stability, especially given the low-scoring nature of football matches and to prevent explosion, the log-scale is used for each weight.

$$log\_w\_statistic \sim N(\mu = log(league\_average), \sigma = 0.5)$$

0.8 Goal Scoring Model

As alluded to earlier, the main novelty is in the Weibull hazard function to model goal scoring patterns 2:

$$H(t = 90; \alpha, \beta) = (\frac{t = 90}{\beta})^{\alpha}$$

where t represents the number of minutes in a match, α is a shape parameter for scoring patterns, and β represents the time between goals scored. The parameter α uses a tight and informative prior using empirical data from Premier League matches 1, where it is observed that the average first goal scored occurs roughly a third of the game through, with subsequent goals occurring at a nonlinear cadence.

$$log\_\alpha \sim N(\mu = log(5), \sigma = 0.1)$$

Similarly, the parameter β is represented by

$$\beta = 90(1 - min(normalized\_max\_time, strength\_ratio))$$

where, again, 90 represents the total number of minutes in a match, $normalized\_max\_time$ (derived from the same empirical data 1) is equal to $\frac{68}{90}$ and signifies that most goals occur before the 68th minute, and $strength\_ratio$ is simply the ratio of one team's attack strength (calculated previously) compared to the opposing team's defense strength. Better teams, such as Manchester City or Liverpool, should have a lower β value and thus fewer times between goals scored.

0.9 Bayesian Inference

PyMC is utilized for Bayesian inference with a sample of $10,000$ and a burn-in of $1,000$ traces (for predicting a single match). The final goal predictions are modeled using a Poisson distribution with the expected goals (xG) as the rate parameter.

$$Goals \sim Poi(\mu = xG)$$

Experimental Setup

Results

As of the time of writing this, there is a huge game this weekend (December 1, 2024): Liverpool vs. Manchester City! Based on the outlined metrics and methodology above, what are some results we can expect (5)? What is the probability of those results? How do the various weights affect the outcome (2 and 3)?

0.10 Offensive Metrics

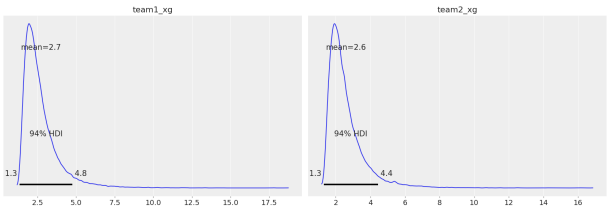

Figure 1: Expected goal distributions for $team\_1$, Liverpool, and $team\_2$, Manchester City

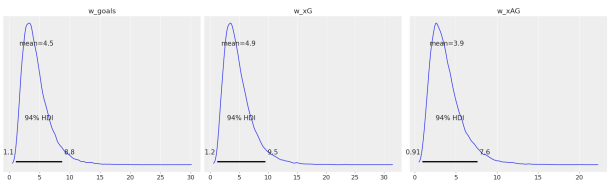

Figure 2: Weighted distribution for goals scored, expected goals, and expected assisted goals

0.11 Defensive Metrics

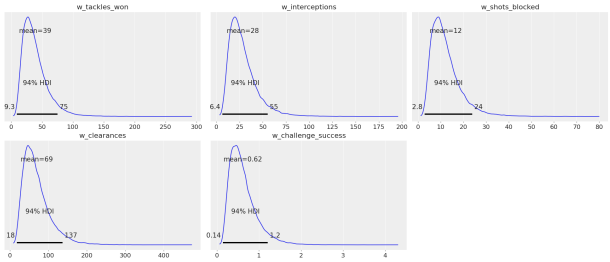

Figure 3: Weighted distributions for tackles won, interceptions, shots blocked, clearances, and challenge success

0.12 Goals Scored

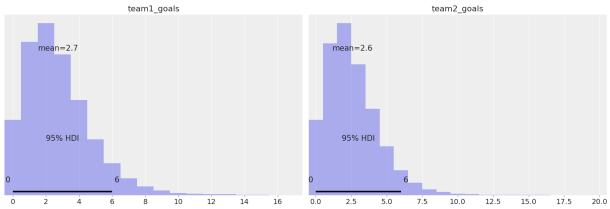

Figure 4: Posterior distributions from the trace of goals scored for each team

0.13 Match Prediction by Probability Distributions

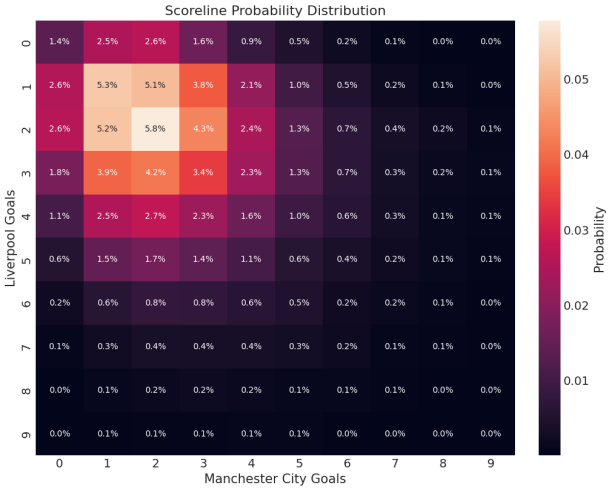

Figure 5: Heat map of scoreline probability distributions for each team: the most common scoreline in the trace is $2-2$, with a 5.8% chance of this occurring

By counting the number of scorelines in the posterior where Liverpool's goals scored exceeds those of Manchester City, there is a 42% chance that Liverpool win, a 40% chance that Manchester City win, and a 18% chance of a draw.