Goal Prediction Model

This project develops a Poisson-based machine learning model to predict the total number of goals in League 1 and League 2 football matches. Using the provided dataset (league_one_and_two_data.csv), Poisson regression and XGBoost models are trained to produce the most accurate distributions of goal probabilities possible. Explanation of the details and rationale behind the modelling process is provided in the Python notebook: model_development.ipynb which is accessible through the GitHub repository.

1) What other data would be most useful to improve the model (outside of simply adding more of the same type of data from other leagues)?

Based on the structure of the Poisson regression model I used, additional data could improve the model by:

i) Creating a more accurate estimation of team attacking and defensive strength for upcoming games:

My model tunes features based on past goals and xG for and against to build estimates of team-specific attacking and defensive strength. These ratings could be improved by more detailed match-level data. For example, shot totals or the number of final-third entries could provide a clearer signal of team capabilities, especially over a short window of games.

Strength ratings could also be improved through a more granular approach. Individual player ratings for teams would allow for a more detailed, team-specific modelling approach. This could use the quality, form and playing style of individual players within a squad to determine a team's tendency to score and concede goals.

Potentially the biggest lever for improving team strength ratings would be financial data on squad/player valuations or wages. Across leagues, team performance is highly correlated with wage expenditure. Having data on how much each team spends on their forwards and defenders would be a very useful proxy for attacking and defensive strength. Financial data of this type would be especially useful for adjusting team ratings at the start of new seasons to account for large shifts in team strengths, which can be created following the transfer window.

ii) Providing more detailed match-specific context:

The provided dataset does not provide much information to help contextualise games. Aside from team strength and home advantage, outcomes of football matches can be significantly affected by: fixture importance/motivation, fixture congestion, injury/suspensions and tactical match-ups. Detailed data reflecting these match-specific contexts could significantly improve the explanatory power of the model.

2) How would you evaluate the model you have built, and what do you see as particular weaknesses?

The model is evaluated primarily using out-of-sample season-by-season cross-validated match-level log-loss. Each season is predicted only using information available prior to that season, which prevents look-ahead bias, and maintains the temporal structure of the data. Full distribution calibration and sharpness are also assessed to ensure predicted goal probabilities align with realised frequencies and that the model does not just regressing towards the league mean.

Overall, the model performs only marginally better than a naïve baseline. This reflects the relatively sparse information set used. The model relies primarily on home advantage exponentially-weighted moving average (EWMA) based attacking and defensive strength, with limited contextual variables. In this context, the modest improvement over chance is unsurprising.

Despite the modest performance, I think my modelling approach carries some merit. The separation of attacking and defensive strength, estimated via tuned EWMA decay parameters, provides a coherent and stable framework which considers long-term quality as well as short-term form. Calibration checks show that total goals are distributed in broadly realistic proportions, indicating that the probabilistic framework is solid.

A weakness of my model is the inability to generate clear signal from promotion and relegation effects. I model using simplified priors and decay terms, which distort early-season strength estimates for transitioning teams. This approach is intuitive theoretically but fails to improve predictive power. In a general sense, the model is overly conservative in its predictions and predicts most games close to the global mean. It struggles particularly with high-scoring matches - underpredicting their frequency and performing poorly when they occur.

3) How would you look to extend/improve the model if you had access to data from multiple other leagues?

Access to data from multiple leagues would strengthen the model by improving estimation precision and structural design.

First, a larger cross-league dataset would improve confidence in the estimated coefficients. The current model is trained within a single competitive environment, which limits variation in match contexts and increases sensitivity to season-specific noise. With more observations across different leagues, the attacking and defensive strength parameters, congestion effects and seasonal dynamics could be estimated more precisely. This would reduce overfitting to league-specific patterns and improve generalisability.

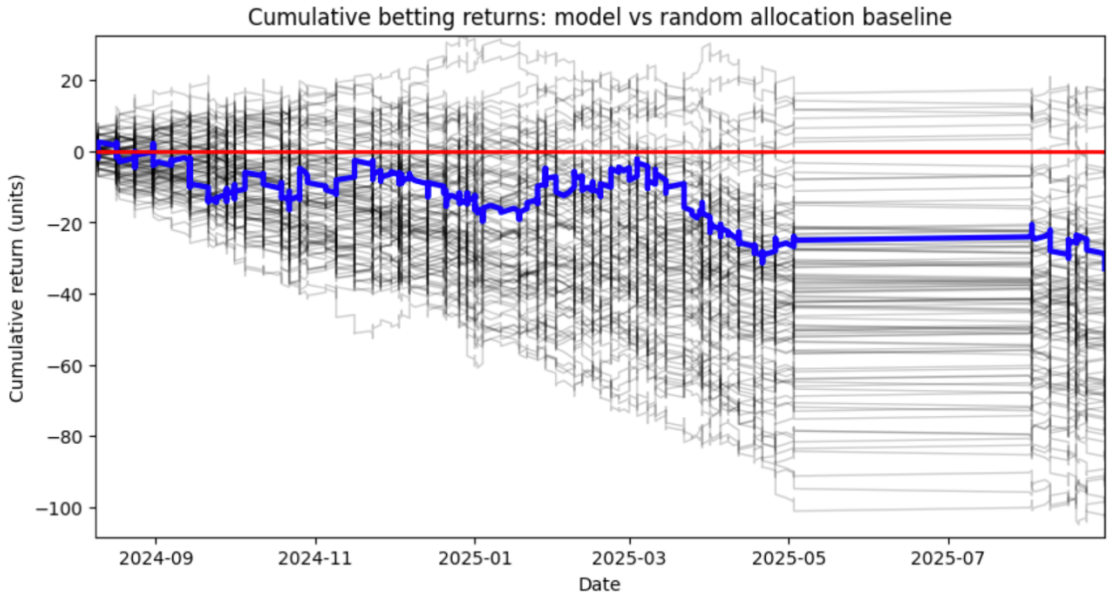
Second, promotion and relegation effects could be modelled empirically rather than heuristically. At present, newly promoted or relegated teams are assigned adjusted priors based on group averages and decay parameters. With multi-league data, it would be possible to estimate the typical attacking and defensive strength differential between tiers. This would allow for data-driven strength shifts when teams change divisions, improving early-season predictions and reducing distortions caused by simplified adjustments.

Finally, cross-league data would allow explicit modelling of structural differences in goal-scoring environments. Different leagues have systematically different baseline scoring rates and variance. Incorporating league-level intercepts or a hierarchical structure would allow the model to account for these differences while still borrowing strength across competitions. This would improve calibration and ensure that predictions remain realistic when applied across varied goal environments.

4) How confident would you feel betting using this model?

Based on the model's performance on the held-out validation set, I would not be too confident betting using this model. With a discrimination score of 0.1, the model is not much better at identifying high or low-scoring games than random chance. To achieve consistent returns, accounting for the book-maker's margin, the model would be required to produce much stronger predictions of future goals than it currently does.

To substantiate this response, I simulated the model's performance using a simple level-stakes betting strategy. Using the model's predictions on the held-out test set, value bets were identified in instances where the model predicted a greater chance of over/under 2.5 goals than the implied odds offered by the average bookmaker. When a stake of one unit was placed on all value bets over the test set, the model produced a cumulative return of -32.96.



To put that into perspective, the chart above compares the performance of the model to a random assignment model, which bets on over 2.5 goals at a rate of P(total goals>2.5). The blue line represents the performance of my Poisson regression model, the red line represents the break-even point, and the grey traces represent 100 simulations of the random-assignment model. The Poisson regression model outperformed the average random-assignment model by 27%. Therefore, this simulation confirms my intuition that whilst my model is better than random chance, I would not be confident betting using it.