Football Corners Challenge

Rishindra Melanathuru

Introduction

The dataset contained 23,830 football matches in the training set (the challenge description said 23,380, so likely a typo) and 341 matches in the test set, each with 9 provided features. There were 13 leagues in total and 333 unique teams. A quick data audit found 8 matches (just 0.03% of the total) with missing home and/or away goals and I removed them rather than impute, although imputing from recent performance would have been possible.

Exploring league-by-league averages revealed a clear home advantage in both goals and corners, with only minor variation across leagues. When plotting matches chronologically, I found six large gaps of over 60 days, typically between November and February. This suggests most leagues here run from February/March to October/November, similar to competition calendars in Scandinavia, parts of Asia, and some African leagues. These gaps split the data into seven seasons: the first six seasons make up the training set, and the final season is the test set.

To ensure feature engineering respected the chronological order, I concatenated the training and test sets, tagging each row as train or test. This allowed me to compute cumulative, rolling, league-wide, and head-to-head features that carried forward naturally into the test period without leaking information from the future. After building these features, shifted by one match to avoid lookahead bias, I fitted and validated models to predict the mean number of total corners for each match, then converted these predictions into under/at/over probabilities. Finally, I used the supplied Asian lines and decimal odds to create a betting strategy, staking from a 341-unit bankroll using the Kelly criterion with push outcomes handled explicitly.

1. Model Description

Before choosing a model, I first explored the distribution of total corners. I plotted the per-match total corners for each league in the training set and also for the full dataset. Across leagues, the variance-to-mean ratio of total corners ranged from about 1.09 to 1.24. Since a Poisson distribution has variance equal to its mean, these ratios being close to one suggested that a Poisson model could be a reasonable starting point.

However, a box plot of all games showed a fair number of outliers, and the all-games histogram had a slightly heavier right tail than a pure Poisson. To account for potential overdispersion and heavier tails, I decided to try both

- **Poisson regression**: assumes that the total number of corners Y_i in match i follows:

$$Y_i \sim \text{Poisson}(\lambda_i), \quad \lambda_i > 0$$

- **Negative binomial regression**: generalises the Poisson by introducing a dispersion parameter $\alpha > 0$:

$$Y_i \sim \text{NB}(\mu_i, \alpha), \quad \text{Var}(Y_i) = \mu_i + \alpha\mu_i^2$$

Features Used

The features were designed to capture both long-term team strength and short-term form, while avoiding any look-ahead bias. All expanding and rolling statistics were **shifted by one match** so that only past information was available when making a prediction for a given match.

In total, 50 features were generated, grouped into the following categories:

1. **Team performance : cumulative (expanding means)** For each team, within each league-season, cumulative averages up to the previous match for Home goals scored / conceded, Home corners for / against, Away goals scored / conceded, Away corners for / against.

2. **Team performance : recent form (rolling means)** Rolling averages over the last 5 matches for the same eight metrics above, capturing short-term trends.

3. **League-wide context : cumulative averages** Expanding means for home/away goals and corners at the league level, both within-season and across all seasons, to capture differences in playing style and tempo between leagues.

4. **Head-to-head (H2H) statistics** Historical averages for goals and corners in prior matchups between the same two teams, separated by home/away context. Also included were prior win/draw/loss rates in this matchup.

5. **Points-based features** Cumulative average points per game for each team, within-season, as a general strength indicator.

6. **Recent results** Rolling win/draw/loss rates for each team over the last 5 matches.

7. **Metadata** Day of week (encoded as an integer) to account for possible scheduling effects.

Handling Missing Values and Feature Refinement

Some matches at the start of a season had missing values for features like cumulative averages (no prior matches to draw from). I imputed these with the training-set mean for each feature; future work could refine this with context-specific methods.

To reduce redundancy, I computed pairwise feature correlations and removed variables with correlations above 0.8, keeping the most interpretable or relevant ones. This cut the set to 40 features. Given the still-wide set and residual multicollinearity, I applied L^2 regularisation in the Poisson regression to control variance, reduce overfitting, and stabilise coefficient estimates (regularisation is discussed later).

Model Fitting, Validation, and Betting Strategy

Validation setup: To simulate predicting on a truly "future" season without data leakage, I split the original training set into: (i) all but the most recent season per league for training, and (ii) the most recent season per league for validation.

Models were compared using Log-likelihood, RMSE between predicted and actual corners, Akaike Information Criterion (AIC).

The Poisson GLM outperformed the Negative Binomial in terms of log-likelihood and AIC, while the RMSE was nearly identical between the two models. Given the slight overdispersion in the data, both were plausible candidates, but the Poisson fit offered a better likelihood-based trade-off.

Probability computation: After fitting on the full training data, the Poisson model produced a predicted mean number of corners (λ) for each match in the test set. Using this, probabilities were computed as:

$$P_{\text{under}} = \sum_{k=0}^{L-1} \text{Pois}(k;\lambda), \quad P_{\text{at}} = \text{Pois}(L;\lambda), \quad P_{\text{over}} = 1 - P_{\text{under}} - P_{\text{at}},$$

where L is the given Asian line.

Betting strategy: With 341 units to stake across the test set, I computed the expected value (EV) for each side, explicitly accounting for possible push outcomes:

$$\begin{aligned} \text{EV}_{\text{under}} &= P_{\text{under}} \cdot (O_{\text{under}} - 1) - [1 - P_{\text{under}} - P_{\text{at}}], \\ \text{EV}_{\text{over}} &= P_{\text{over}} \cdot (O_{\text{over}} - 1) - [1 - P_{\text{over}} - P_{\text{at}}], \end{aligned}$$

where P_{under}, P_{at}, and P_{over} are the model-predicted probabilities and O denotes the decimal odds.

Stakes were then sized using a Kelly criterion variant that incorporates pushes:

$$f^* = \frac{P_{\text{win}} \cdot b - [1 - P_{\text{win}} - P_{\text{push}}]}{b}, \quad b = O - 1,$$

clipped to $[0, 1]$ and then scaled to respect the total bankroll.

The unpenalised Poisson GLM achieved the highest backtested ROI (0.51) and Sharpe ratio (1.78), but has no mechanism to control overfitting, a concern given the relatively wide, correlated feature set. To address this, I also fitted `scikit-learn`'s `PoissonRegressor` with L^2 regularisation, standardising all features and tuning α (the regularisation strength) via validation to 5.455. The regularised model's ROI (0.14) and Sharpe ratio (1.19) were lower on this dataset, but the penalty makes it more robust to noise and less likely to chase spurious correlations, particularly when applied to entirely new seasons.

Both ROI and Sharpe ratio were computed as:

$$\text{ROI} = \frac{\text{Net profit}}{\text{Total staked}}, \quad \text{Sharpe ratio} = \frac{\mathbb{E}[R]}{\sigma_R}, \tag{1}$$

where R is the per-bet return.

Before this challenge, I had been working on a related IPL cricket win-probability model, motivated by the limitations of ESPNcricinfo's "WinViz." I created custom features (batter score, bowler score, head-to-head index, pressure index) and began exploring a jump-diffusion framework, inspired by my PhD research on quantum measurement under noise.