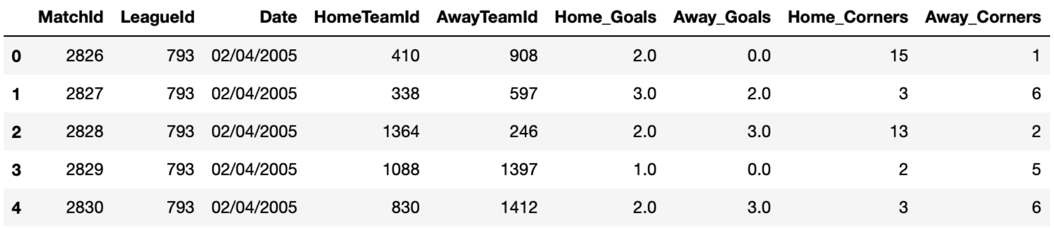
Corners Challenge

1 Introduction

In response to the challenge, I have developed a mathematical model for determining the distribution of total corners in football matches using a comprehensive training set. This model assumes a certain distribution for corners, leveraging features such as historical corner data for the home and away teams as well as historical league specific corner data. For each match in the provided test set, my model generates probabilities for corners being under, at, or over the given line. I also devise a simple strategy as to whether to bet on a match, the selection to choose, and the suggested stake amount.

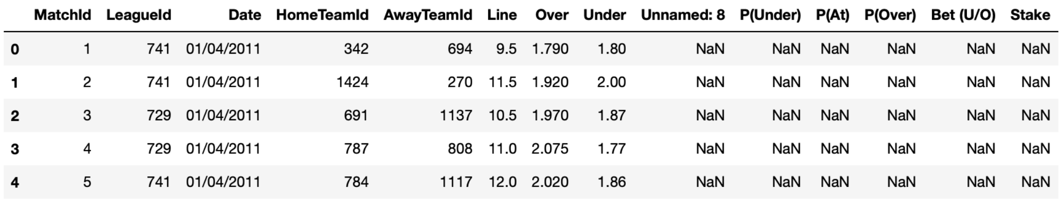
2 Exploratory Data Analysis



	MatchId	LeagueId	Date	HomeTeamId	AwayTeamId	Home_Goals	Away_Goals	Home_Corners	Away_Corners
0	2826	793	02/04/2005	410	908	2.0	0.0	15	1
1	2827	793	02/04/2005	338	597	3.0	2.0	3	6
2	2828	793	02/04/2005	1364	246	2.0	3.0	13	2
3	2829	793	02/04/2005	1088	1397	1.0	0.0	2	5
4	2830	793	02/04/2005	830	1412	2.0	3.0	3	6

Figure 1: First five rows of the training dataset

In the process of building models, diving into the data through exploratory analysis is like turning on the lights in a dark room. It helps us understand the story behind the numbers, spot interesting trends, and decide how to build our models effectively. The training data provided in the 'train.csv' file records the number of goals and corners recorded by the home and away teams in various football leagues between 2005 and 2010. The 'HomeTeamID', 'AwayTeamId', 'LeagueId' and 'MatchId' columns relate each match to the league to which it belongs as well as the home and away teams competing. Dates of the fixtures are also provided. The test dataset provides details of the teams and league of football matches in a two month window between 01/04/2011 and 23/05/2011, however goals and corner data are omitted. Instead, the test set includes odds and lines in order to devise the betting strategy.



	MatchId	LeagueId	Date	HomeTeamId	AwayTeamId	Line	Over	Under	Unnamed: 8	P(Under)	P(At)	P(Over)	Bet (U/O)	Stake
0	1	741	01/04/2011	342	694	9.5	1.790	1.80	NaN	NaN	NaN	NaN	NaN	NaN
1	2	741	01/04/2011	1424	270	11.5	1.920	2.00	NaN	NaN	NaN	NaN	NaN	NaN
2	3	729	01/04/2011	691	1137	10.5	1.970	1.87	NaN	NaN	NaN	NaN	NaN	NaN
3	4	729	01/04/2011	787	808	11.0	2.075	1.77	NaN	NaN	NaN	NaN	NaN	NaN
4	5	741	01/04/2011	784	1117	12.0	2.020	1.86	NaN	NaN	NaN	NaN	NaN	NaN

Figure 2: First five rows of the test dataset

Let's look at the distributions of corners for the home and away teams. Observe in Fig. 3 that the distribution of away team corners is narrower and slightly to the left of home team corners. It has a smaller mean and a smaller variance. Hence, there is more predictability in the distribution of away team corners and the home teams tend to win a greater number of corners in a given match.

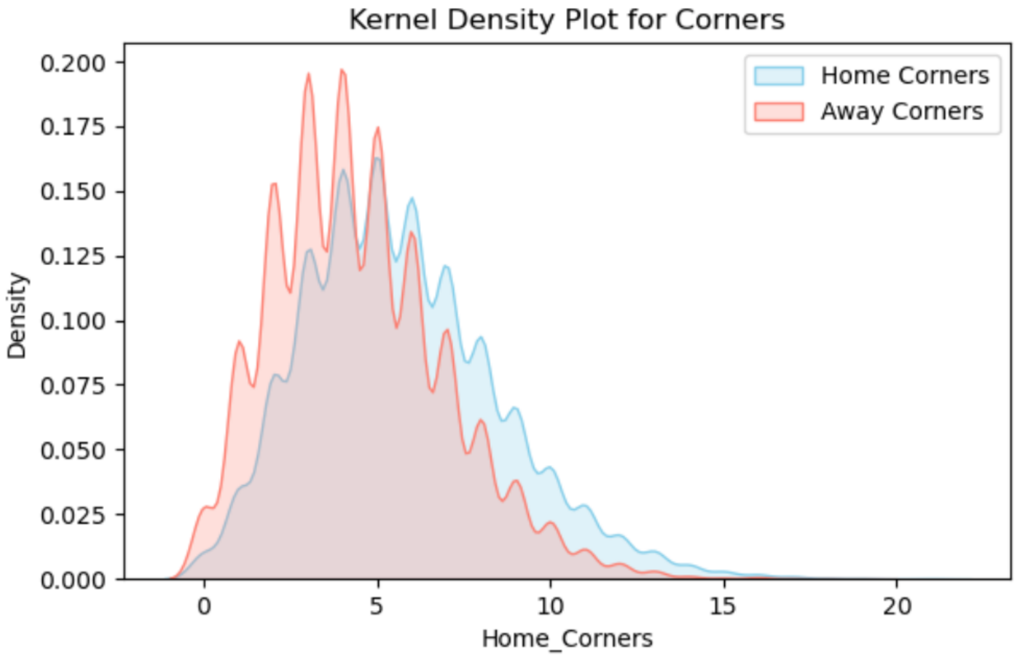


Figure 3: Kernel density plots of home and away team corners for the training data

Now, the purpose of this challenge is to estimate the distribution of total corners in a given match. It is highlighted in Table 1 that the most common number of total corners in a match is 9 despite the mean and median being 10.22 and 10, respectively. A variance-to-mean ratio of 1.17 for the entire dataset indicates a slightly higher level of variability compared to a perfectly Poisson-distributed dataset (which would have a ratio of 1). Hence, there appears to be some degree of overdispersion in the data.

Statistical Measure	**Value**
Mean	10.22
Median	10.0
Mode	9
Q1	8.0
Q3	12.0
Standard Deviation	3.45
Variance	11.91
Variance-to-Mean Ratio	1.17

Table 1: Overview of descriptive statistics for total corners in the training set

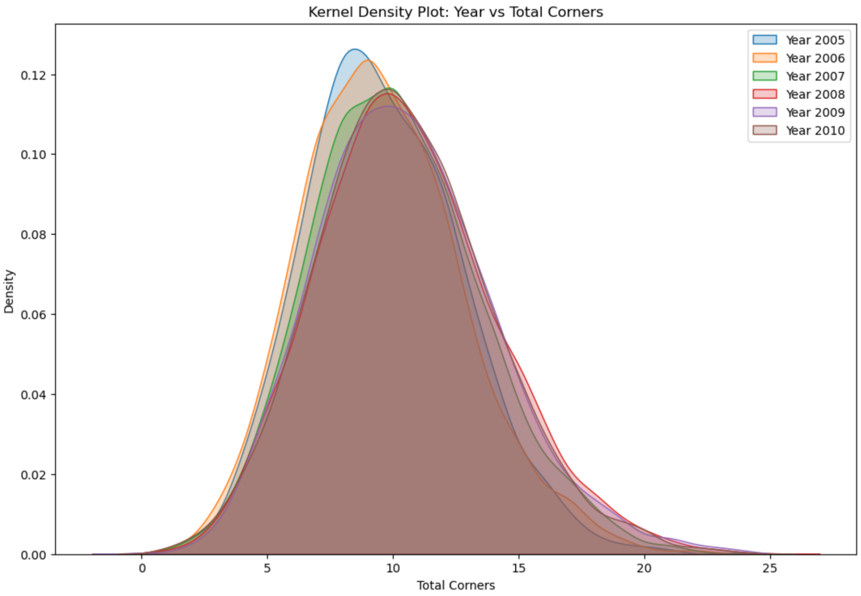


Figure 4: Kernel density plots of total match corners for each different year present in the training data

The training data includes matches from a six-year period, hence it may be useful to compare the distributions of total match corners for each of these years. Observe in Fig. 4. that the distributions of total corners in 2008, 2009 and 2010 are very similar while the distributions for 2005 and 2006 are narrower and shifted slightly to the left. The 2007 distribution is somewhere in between these two clusters. I conducted pairwise Kolmogorov-Smirnov tests on the samples from the years 2005 to 2010 to assess the similarity of their distributions. The results revealed that the samples from 2008, 2009, and 2010 do not exhibit a statistically significant difference from each other at a significance level of 0.05. This suggests that the distributions of data for these three years are comparable. However, the distributions of data from the remaining years (2005, 2006, and 2007) are found to be statistically significantly different from those of 2008, 2009, and 2010. This implies a notable shift or change in the data distribution during these years. Given these findings, and that the task is to predict total corners for 2011 data, I made the decision to focus the analysis on data from the three year period of 2008-2010. This decision is based on the observed stability and similarity in the distribution of these three years' data, providing a more consistent foundation for predictive modeling.

Professional football leagues around the world are characterised by their subtle variations in tactics and style of play. Hence, it is reasonable to assume that the distributions of match corners may differ slightly depending on the league in question. In Table 2., these subtle differences are illustrated. The mean number of total corners ranges from 9.65 to 11.24. While the differing variance-to-mean ratios across leagues highlight diverse data distributions. Understanding this helps tailor models for more accurate predictions, acknowledging the unique patterns within the dataset.

LeagueId	**Mean**	**Median**	**Mode**	**Variance**	**Variance-to-Mean Ratio**
729	11.24	11.0	10	14.29	1.27
734	9.86	10.0	9	10.66	1.08
741	10.98	11.0	10	12.60	1.15
764	10.63	10.0	10	12.85	1.21
776	9.85	10.0	10	10.54	1.07
781	10.70	10.0	9	13.22	1.24
793	11.23	11.0	12	12.68	1.13
795	9.65	9.0	9	11.37	1.18
800	10.44	10.0	11	11.98	1.15
801	10.49	10.0	10	12.41	1.18
811	9.74	9.0	8	12.11	1.24
813	10.06	10.0	9	11.40	1.13
816	10.97	11.0	10	12.65	1.15

Table 2: Comparison of data distribution for the different leagues present in the training data

3 Feature Engineering

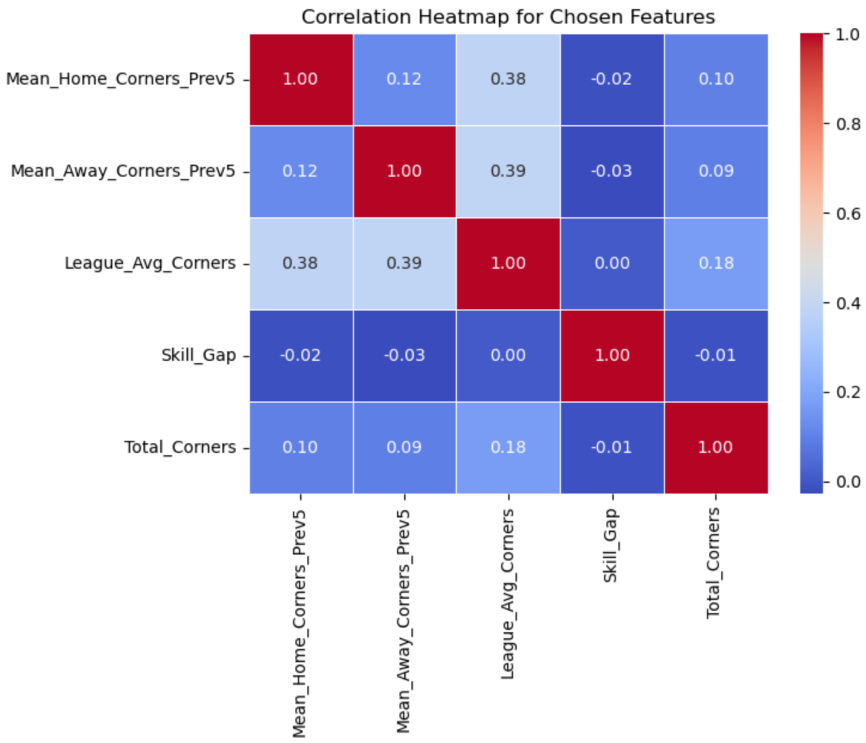


Figure 5: Correlation heatmap of the features selected for modeling

The difference in the skill level of the two teams competing may have some correlation with the number of corners in a match. With this in mind, I used the match results and goal differences to create an ELO rating system for the teams in the training data, giving each team an initial rating of 1500 and updating after every match. More information about this rating system can be found at [1]. It is reasonable to assume that a teams recent history of match corners may be an important explanatory variable. Hence the mean number of total corners in the home team's previous five home matches as well as the mean number of total corners in the away team's previous five away matches are calculated. Teams can often vary their playing style based on whether they are playing at home or away so this is the reason why I only considered the last five home games of the home team and vice versa. The significance of the league to which the match belongs was outlined in Section 2. For this reason, a rolling average number of corners is incorporated into the analysis for each of the leagues with a minimum window size of 50 to ensure a large enough sample size. Observe in Fig. 5, the league average corners feature as well as

[1] `https://www.eloratings.net/about`

the historical corner data for the home and away teams exhibit small positive correlation with the total number of corners. However, the 'Skill Gap' feature defined as the absolute value of the difference in rating of the two teams (accounting for home advantage) doesn't seem to be correlated very well with the number of Total Corners. For this reason and for the sake of simplicity, this feature will dropped before fitting the models in the next section.

4 Model Fitting

I employed three different statistical models - Poisson, Geometric Poisson, and Negative Binomial - in order to predict the distribution of total corners, incorporating the features outlined in Section 3. The Geometric Poisson and Poisson models assume a specific distribution, while the Negative Binomial model accounts for overdispersion. Despite slight overdispersion in the training data, the choice of the Naive Poisson model is justified as it gives rise to the greatest log-likelihood. Furthermore, its assumption of equal mean and variance aligns reasonably well with the data characteristics.

Table 3: Model Comparison

	Log Likelihood	Residual df	Deviance	Chi-squared
Geometric Poisson Model	-36055	13368	16709	16,700
Poisson Model	**-35533**	13367	15665	15,400
Negative Binomial Model	-45395	13367	1471.2	1,350

5 Betting Strategy

The test data is augmented with the relevant features and the estimated Poisson distribution is employed to generate an estimate for the Poisson parameter, λ, for each sample. The historical corner data for the test samples is imputed with the values from the five most recent samples from the training set. The league average corners per game is calculated for each unique league on the entire training set (from 2008 onwards). Using these estimated parameters and the betting line, estimates for the probability of the number of corners being under, at, and over the given line are generated ($P(Under), P(At), P(Over)$. These are combined with the O/U odds to determine the expected value of betting the over and the expected value of betting the under (accounting for push bets when the line is an integer). Expected values are found by multiplying the probability of each event occurring by the associated gain or value, and then summing these products across all possible events. Hence, the formulae for the expected values $\mathbf{E}(Over)$ and $\mathbf{E}(Under)$ are given by

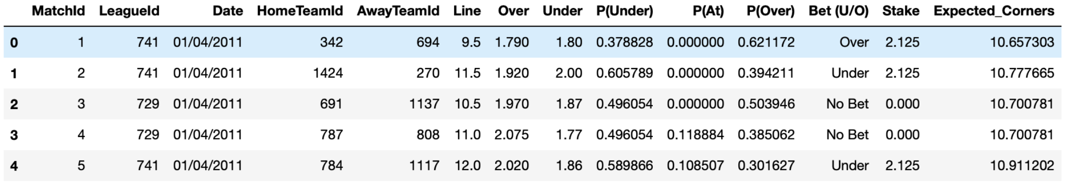
$$\mathbf{E}(Over) = (Odds(Over).P(Over) + P(At)) - 1 \tag{1}$$

$$\mathbf{E}(Under) = (Odds(Under).P(Under) + P(At)) - 1 \tag{2}$$

The expected value $\mathbf{E}(Under)$ represents the anticipated profit for an under bet of 1 unit, while $\mathbf{E}(Over)$ denotes the expected profit for an over bet of 1 unit under the model's assumptions. In theory, with a perfect model, putting all 341 units on the game with the highest positive deviation from zero in expected value would maximise return. However, exploring alternative, potentially less risky strategies may be prudent given uncertainties in real-world football data. Out of the 341 matches in the test set, 251 show potential betting advantages, where either the expected value for 'Over' or 'Under' exceeds zero. This suggests potential profitability in placing bets on these games, as indicated by our model's estimated distributions. Additionally, prioritising games with a more substantial positive deviation from zero signals a higher bias in the odds according to our model. There are 80 games where the expected value of one of the bets exceeds 0.1. A simple strategy will be pursued, betting 1 unit on each match where the expected value of an O/U bet is above 0 and allocating the remaining 90 units equally among the 80 matches with an expected value greater than 0.1 for either the under bet or over bet. Hence, 1 unit will be placed on each of the 171 games with moderately biased odds, and 2.125 units will be placed on each of the 80 games with more substantially biased odds (according to the Poisson model presented in Section 4). There are 90 games on which no bets will be placed. As it turns out, 222.5 units are staked on 'under' bets and the remaining 118.5 units are staked on 'over' bets. This suggests an overall inclination toward favourable odds for placing 'under' bets on corners.

6 Conclusion

In Section 2, we explored the data to uncover connections and enhance our insights. Using features related to the home team's corner history, the away team's corner history, and league-specific corner history, three explanatory variables were devised. Applying statistical models to the training data, the Poisson distribution emerged as the most likely predictor. Probabilities and expected values for Over/Under bets were generated from the estimated parameters for the test data. Based on the expected values, a simple betting strategy was created. If the expected value of a bet is between 0 and 0.1 units, one unit is staked, and if the expected value of a bet is greater than 0.1 units, 2.125 units are staked. With this strategy, greater odds biases perceived by the model are capitalized



	MatchId	LeagueId	Date	HomeTeamId	AwayTeamId	Line	Over	Under	P(Under)	P(At)	P(Over)	Bet (U/O)	Stake	Expected_Corners
0	1	741	01/04/2011	342	694	9.5	1.790	1.80	0.378828	0.000000	0.621172	Over	2.125	10.657303
1	2	741	01/04/2011	1424	270	11.5	1.920	2.00	0.605789	0.000000	0.394211	Under	2.125	10.777665
2	3	729	01/04/2011	691	1137	10.5	1.970	1.87	0.496054	0.000000	0.503946	No Bet	0.000	10.700781
3	4	729	01/04/2011	787	808	11.0	2.075	1.77	0.496054	0.118884	0.385062	No Bet	0.000	10.700781
4	5	741	01/04/2011	784	1117	12.0	2.020	1.86	0.589866	0.108507	0.301627	Under	2.125	10.911202

Figure 6: Test dataset with the probabilities, bet types and stakes added to the relevant columns

on by increasing the stake. The probabilities $P(Under), P(At)$, and $P(Over)$ are added to the test dataset along with the decision on which type of bet to place (if at all) and the stake, all according to the estimates of the fitted model.