Predicting football results

Supervision Meeting 3

Timo Lechner

2025-01-19

Project Setting

In the following we want to analyse football results from the German Bundesliga, beginning with the season 2005/06, up to the season 2023/24. Data is taken from the website https://www.football-data.co.uk/data.php. Specifically, we want to create different models to predict the full time result of the individual games. From the perspective of the home team a game can have three different outcomes: win, draw or loss. The result is given in the column "FTR" (Full Time Result) and is either "H" (home win), "D" (draw) or "A" (away win). Furthermore, the result can be obtained by looking at the goals scored by the home and away team, respectively. This is given by the columns "FTHG" and "FTAG" (Full Time Home/Away Goals).

The Bundesliga consists of 18 teams and one season has 34 matchdays. This leads to 304 games per season. We have data for 19 seasons available and hence 5814 entries in total. Since the bottom two teams of each season get relegated to the second league and two teams from the second league get promoted, we have 36 teams that played at least once in the German Bundesliga in the last 19 seasons.

This format prevents us from trivially splitting our data into a training and a test data set from a specific season on, since training data for some teams might not be available if they got promoted in a more recent season only. Hence, we will include the first few matches of a season as well in our training data to improve our data basis. More specifically, we will use all data until the season 2017/18 as training data. From this season on, games that were played between January and June fall into our test data set. This approximately corresponds to the second half of the season. This way we end up with 4679 entries for our training data and 1135 entries for the test data, which reflects a proportion of 80.5% : 19.5%.

We will develop four different models to predict the outcome (home win, draw, away win) of the individual games. The first one is a classic Dixon-Coles model which includes two independent Poisson distributions together with a factor to capture low-scoring matches and a factor that captures the home team advantage. The second and third model are Bayesian models, following the approach of Baio & Blangiardo, where Bayesian inference is used on Poisson distributions. The second model will have a general home team advantage factor, while the third model has a team-specific home team advantage factor. The German Bundesliga is known for its high attendance in stadiums and historical teams. We can analyse if this effect differs between the clubs and if it has an effect on the quality of the prediction as well. Lastly, we will develop a neural network model using TensorFlow and see how this performs in comparison to the "standard" models. As features we will use the performance of the home/away team in the last 8 games as well as the performance in the whole past.

In order to compare the models we will look at two metrics in detail. The first one is the so called Brier Score. It measures the accuracy of the probabilistic predictions for multi-category forecasts and is defined in the following way:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}(o_{ij}-p_{ij})^2$$

Here, m is the number of possible outcomes/classes (in our case three) and n the number of instances (in our case 1132 number of games in the test data set). o_{ij} describes the binary outcome of the i-th game for the j-th class and p_{ij} the predicted probability of that case.

The second metric we will look at is the Ranked Probability Score. Other than the Brier Score, it takes the distance between classes into consideration, i.e. classes 1 and 2 are considered closer than classes 1 and 3. This makes sense in our setting, since a home win is closer to a draw than an away win. It is defined as:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m-1}\sum_{k=1}^{j}(o_{ij}-p_{ij})^2$$

Dixon-Coles Model

The Dixon-Coles Model incorporates various factors such as team strength, match location (home or away), and historical performance to estimate the probability of different match outcomes. It uses Poisson regression to model the number of goals scored by each team, which assumes that goals are scored randomly but at a predictable average rate. The model includes a specific parameter γ to account for the common phenomenon of home advantage, where teams tend to perform better when playing at their own stadium. It furthermore incorporates an adjustment factor ρ that accounts for low-scoring games. Each team has an offence α and a defense parameter β and the expected number of goals in a game is given by:

$$home : \lambda = \alpha_{home} \cdot \beta_{away} \cdot \gamma$$
$$away : \mu = \alpha_{away} \cdot \beta_{home}$$

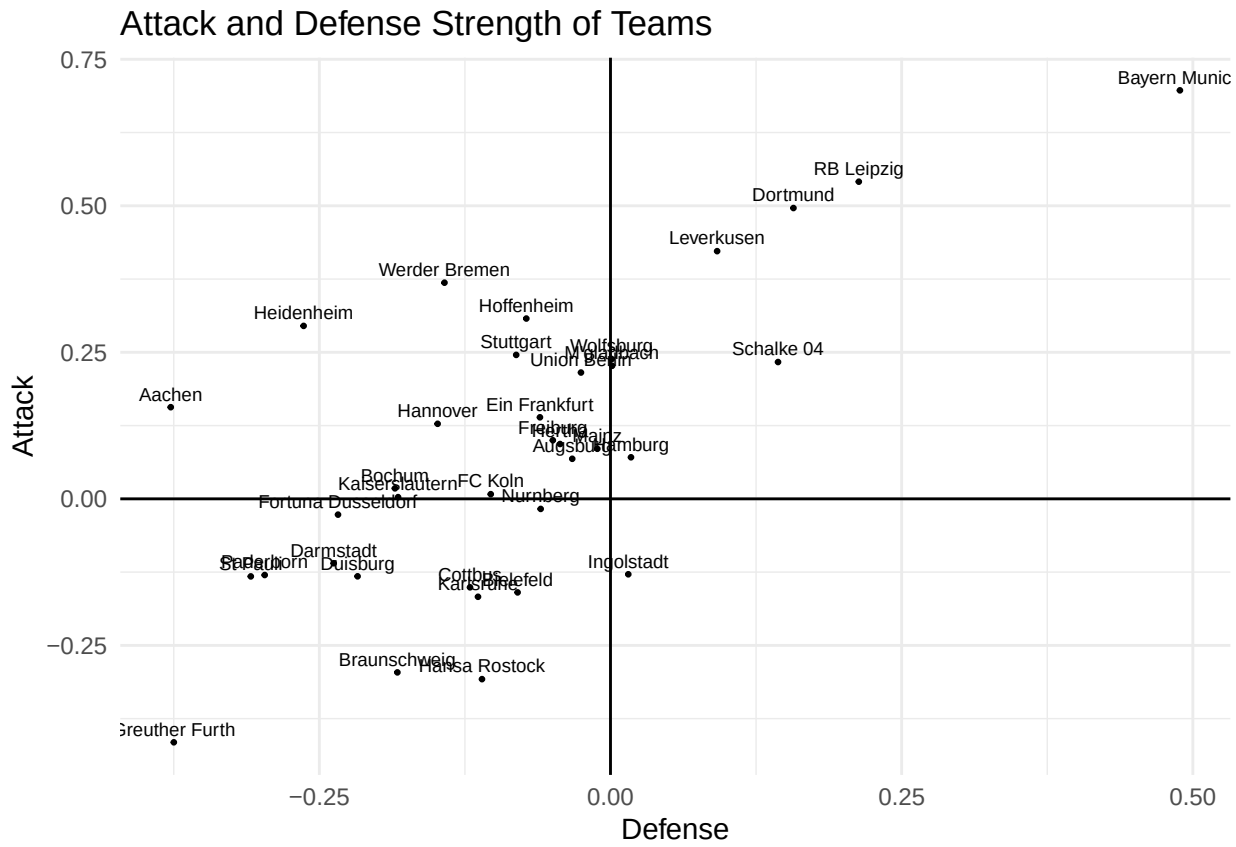
Dixon and Coles found that the standard Poisson model underestimates the frequency of low-scoring games, especially draws. Hence they introduced a bivariate adjustment ρ, which dictates the degree of correlation between the probabilities of low-scoring outcomes. The probability of the outcome is then adjusted by the following factor:

$$adjustment = \begin{cases} 1-(\lambda\cdot\mu\cdot\rho), & x=0,\ y=0 \\ 1+(\mu\cdot\rho), & x=1,\ y=0 \\ 1+(\lambda\cdot\rho), & x=0,\ y=1 \\ 1-\rho, & x=1,\ y=1 \\ 1 & else \end{cases}$$

Here, x is the number of goals scored by the home team and y the number of goals scored by the away team.

Results

First, let us look at the optimized parameters of the attack and defense strength of the individual teams.



The general home team advantage is given by:

```
## home_advantage
##       0.256846
```

The low-scoring adjustment factor is given by:

```
##         rho
## -0.09061691
```

Running this model against our test data we receive the following values for our two metrics for comparison:

```
## [1] 0.6005681
```

```
## [1] 0.4164974
```