Master Thesis

Football score modelling

Timo Lechner
GUID: 2838994L

University of Glasgow
12 March 2025

1 Introduction

2 Data

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 based on historic data. 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 a specific season as training data. From this season on, games that were played between January and June fall into our test data set.

3 Methods

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, where Bayesian inference is used on Poisson distributions. The second model will have a general home team advantage factor and was first described by Baio and Blangiardo, while the third model has a team-specific home team advantage factor as well as a low-score adjustment similarly to the Dixon-Coles model. 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 several game statistics like goals, shots, corners and fouls.

In order to compare the models we will look at the Ranked Probability Score. It measures the accuracy of the probabilistic predictions for multi-category forecasts and is defined in the following way:

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

Here, m is the number of possible outcomes/classes (in our case three) and n the number of instances. 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 Ranked Probability Score 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.

4 Models

4.1 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.

4.2 Baio Blangiardo Model

Next we want to try a Bayesian approach that combines prior knowledge with new data to update beliefs, allowing for the estimation of parameters and making predictions. The prior distribution represents the initial beliefs about the parameters before observing any data. The choice of priors reflects the researcher's beliefs or previous knowledge about the parameters. The Likelihood function models the probability of observing the data given the parameters. It reflects how likely the observed outcomes are, given certain parameter values. The posterior distribution is the updated belief about the parameters after observing the data. It is derived using Bayes' theorem:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

where $P(\theta|D)$ is the posterior distribution, $P(D|\theta)$ is the likelihood, $P(\theta)$ is the prior distribution and $P(D)$ is the marginal likelihood, which is often a normalizing constant.

Again, we are assuming two Poisson distributions for the number of home and away goals and are taking attack, defense and home advantage parameters into account. As new match data is observed, the model updates the estimates for the parameters through the likelihood function. This updating process refines the estimates of the model parameters. Once the posterior distributions of the parameters are obtained, predictions about future match outcomes can be made. This involves simulating the number of goals for each team in future matches based on the estimated parameters.

For the priors we choose $N(0,1)$ distributions. Baio and Blangiardo included additional parameters μ for the mean and τ for the precision following a Gamma distribution in their model. For the sampling we will use 2000 iterations with a warmup phase of 1000 iterations and 4 chains in total.

4.3 Extended Bayesian Model

In the next step we want to extend our Bayesian Model by making the home team advantage parameter dependent on the team. Furthermore, we include the low-score adjustment ρ from the Dixon-Coles model.

4.4 Neural Network Model

We will implement our last model using neural networks. For this reason we need to create multiple features which could help us determine the outcome of a soccer match. We will therefore use the average per game of the following statistics across the whole historic data as well as the current season only:

- Wins
- Draws

- Losses
- Goals scored
- Goals conceded
- Shots taken
- Shots received
- Fouls made
- Corners taken
- Yellow cards received
- Red cards received

For implementation, we will use Keras. We will be running five different models with different architecture and use the season before our prediction season as validation to select the best model based on the Ranked Probability Score. The output layer consists of three units representing the possible outcomes: HomeWin, Draw, and AwayWin, with a softmax activation function to produce probabilities that sum to one. The model is compiled using categorical crossentropy as the loss function, which is suitable for multi-class classification tasks. The Adam optimizer is used with a learning rate of 0.001, and accuracy is used as a metric to evaluate performance. It is trained for 100 epochs, with a batch size of 32.

5 Results

5.1 Dixon-Coles Model

Home Team Advantage	Low-Score Adjustment	RPS	Accuracy
0.277	-0.137	0.410	0.485

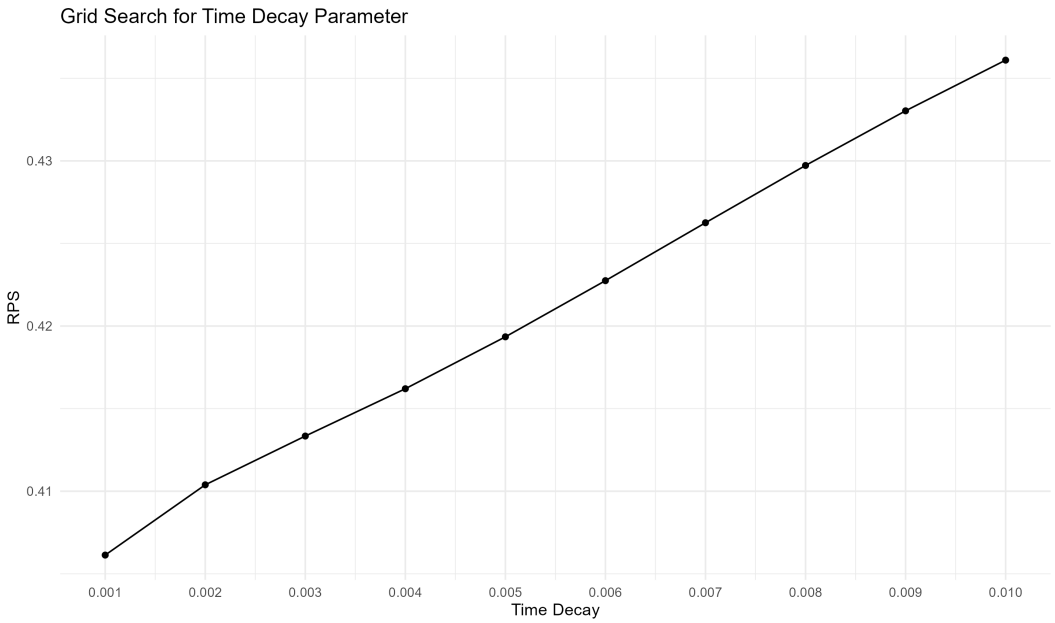

Figure 1: Grid Search for best time decay parameter for Dixon-Coles model

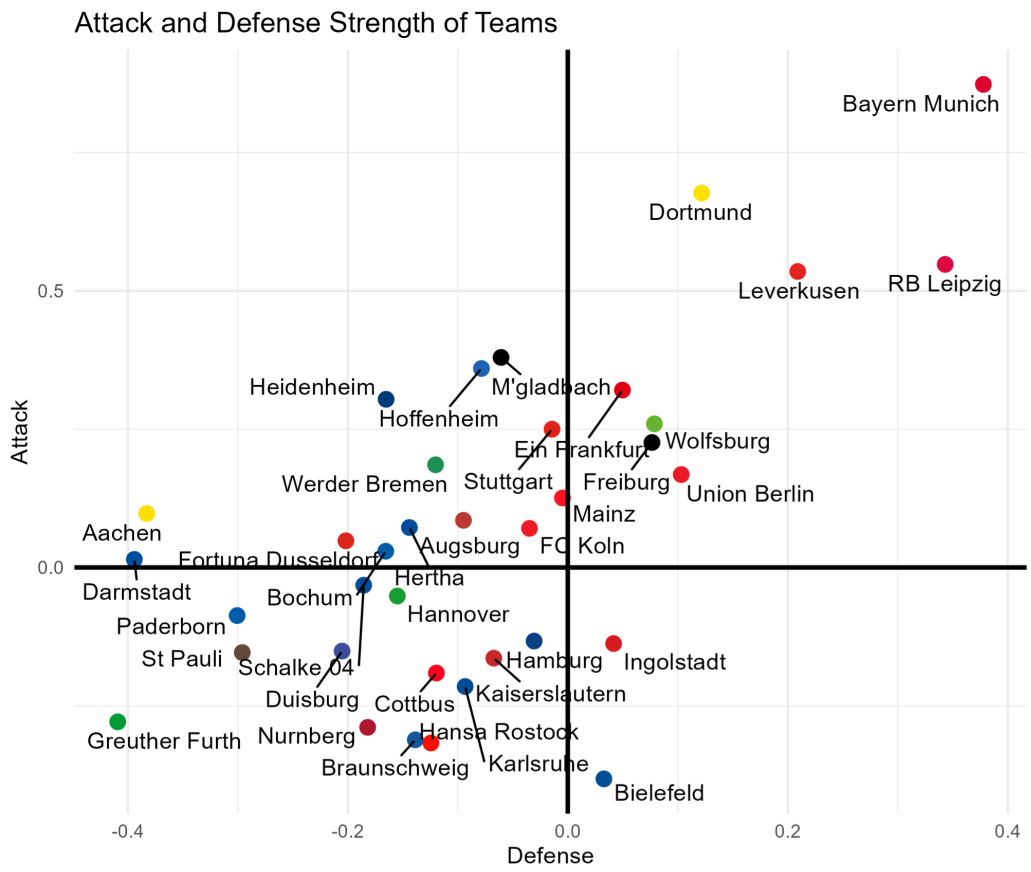

Figure 2: Parameter plot for Dixon-Coles model

5.2 Baio-Blangiardo Model

RPS	Accuracy
0.424	0.503

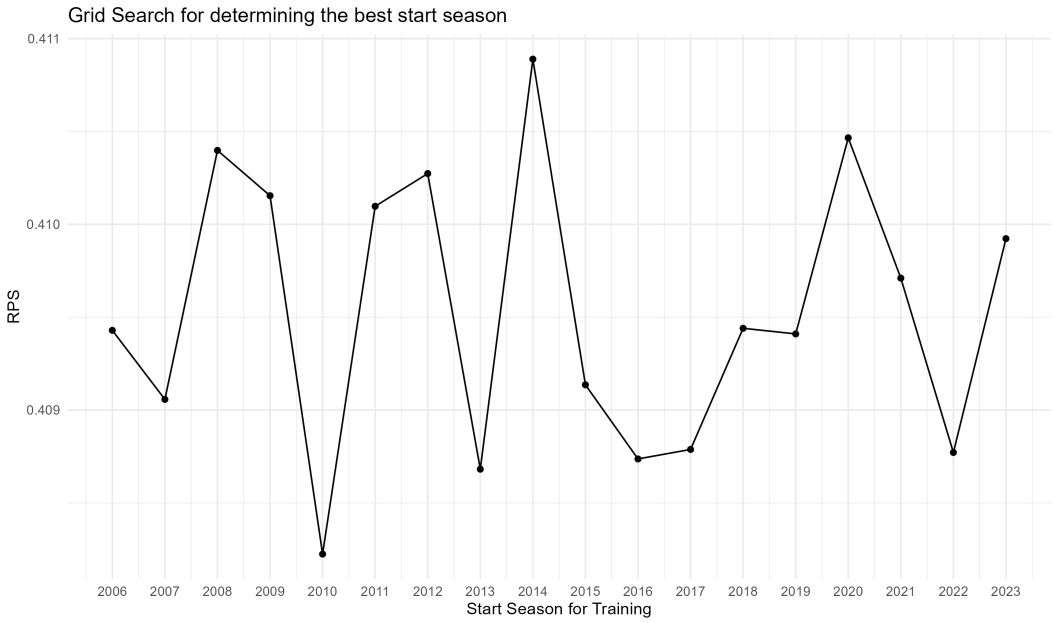

Figure 3: Grid Search for best start season for Baio-Blangiardo model

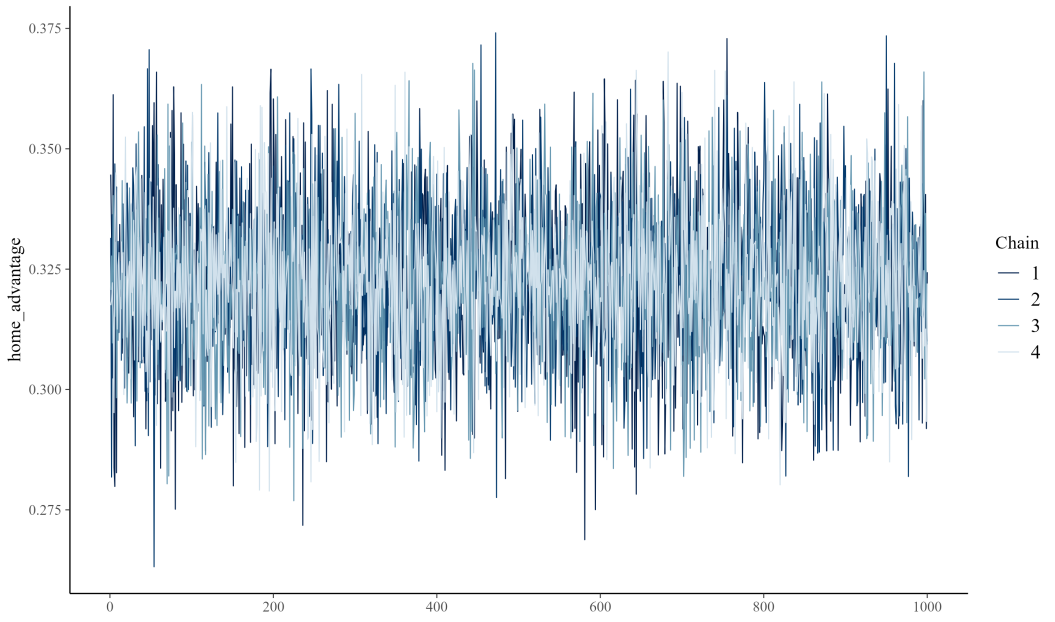

Figure 4: Trace plot for Markov chain Monte Carlo for Baio-Blangiardo model

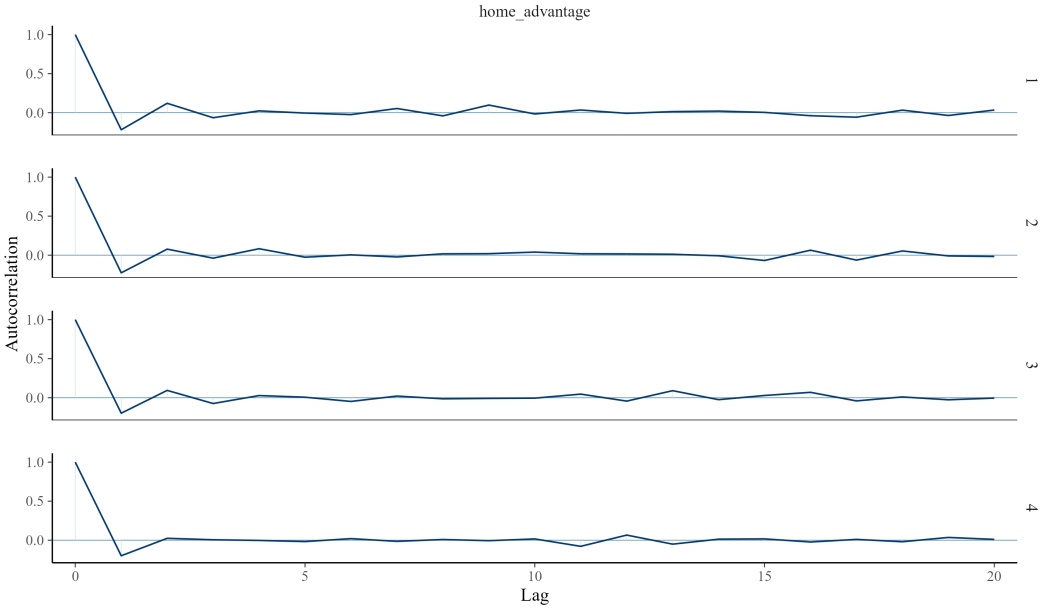

Figure 5: Autocorrelation for Markov chain Monte Carlo for Baio-Blangiardo model

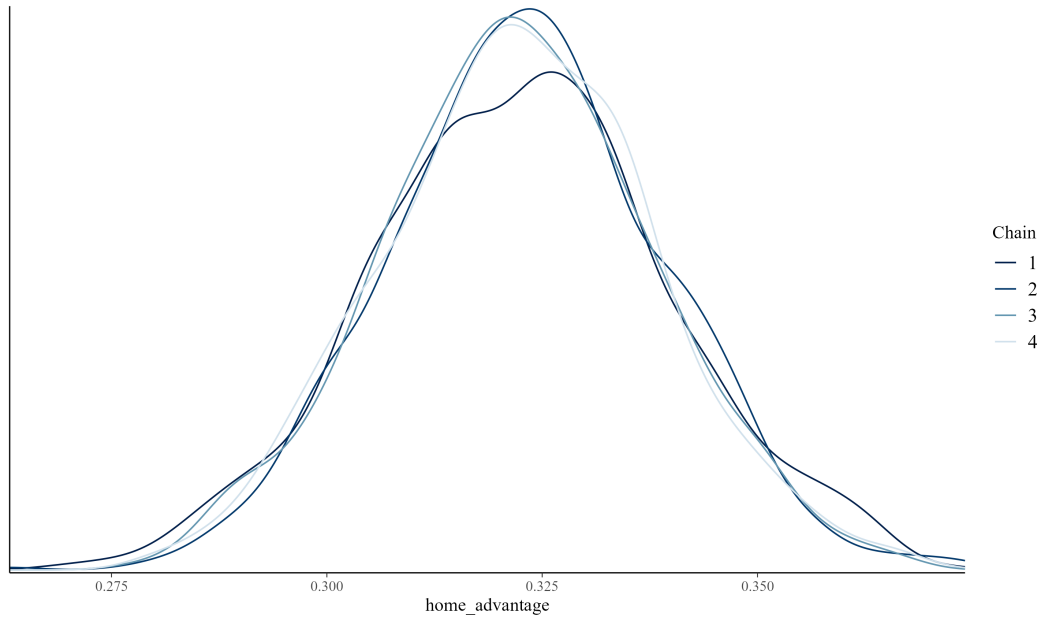

Figure 6: Distribution overlay for Markov chain Monte Carlo for Baio-Blangiardo model

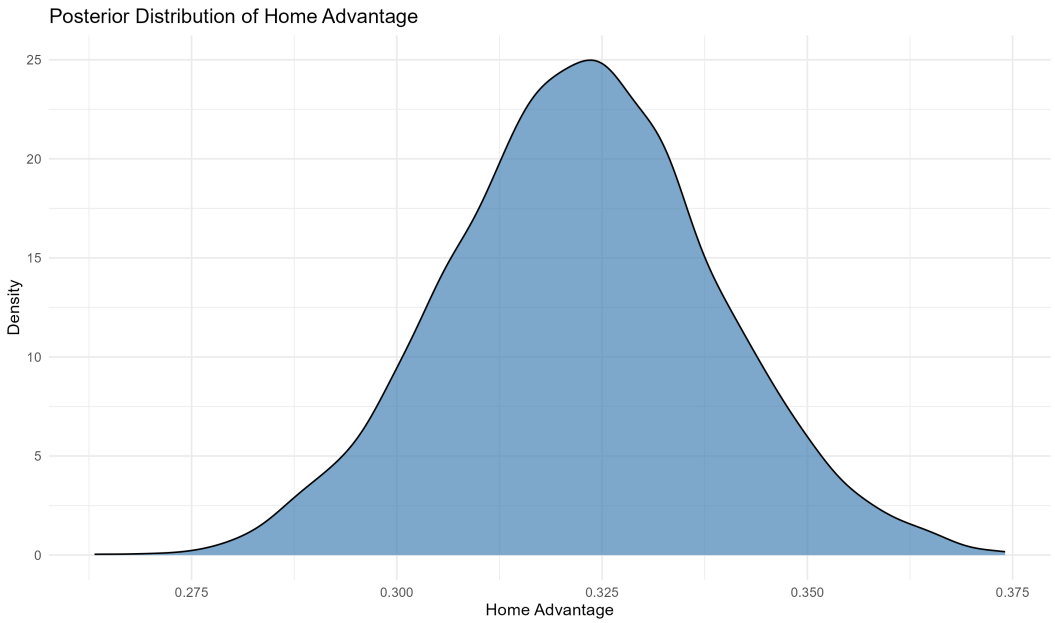

Figure 7: Distribution of home team advantage for Baio-Blangiardo model

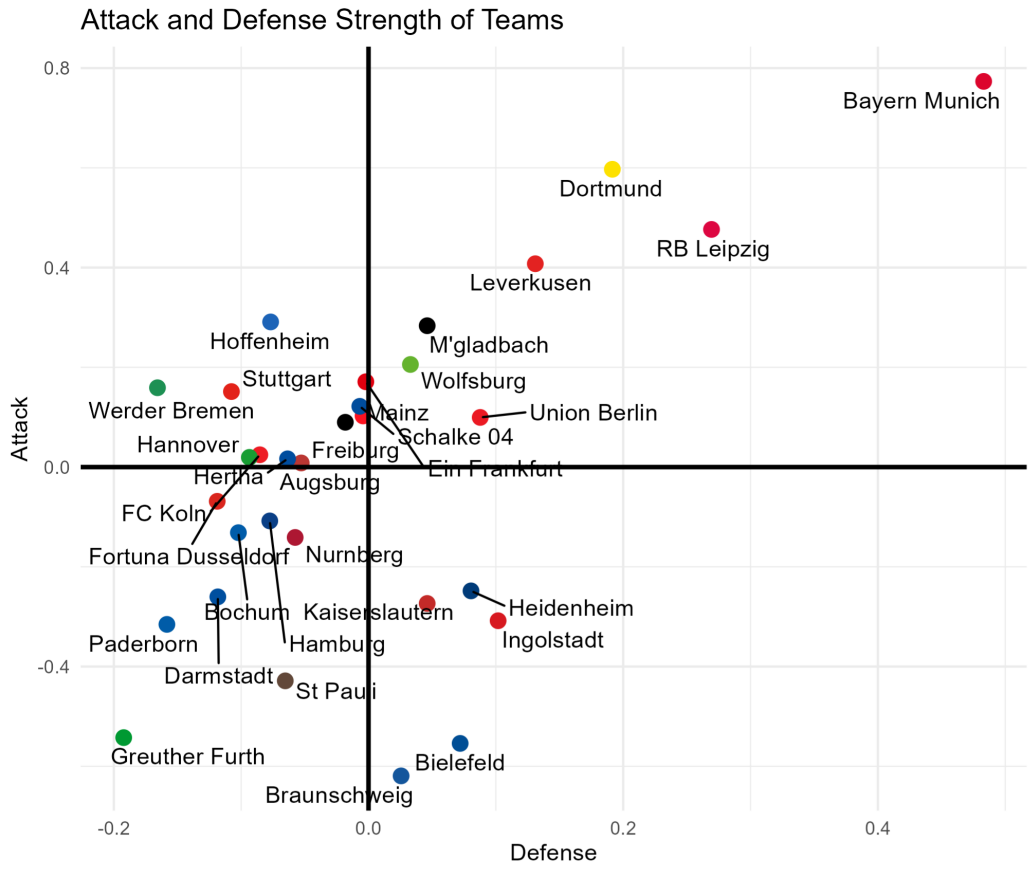

Figure 8: Parameter plot for Baio-Blangiardo model

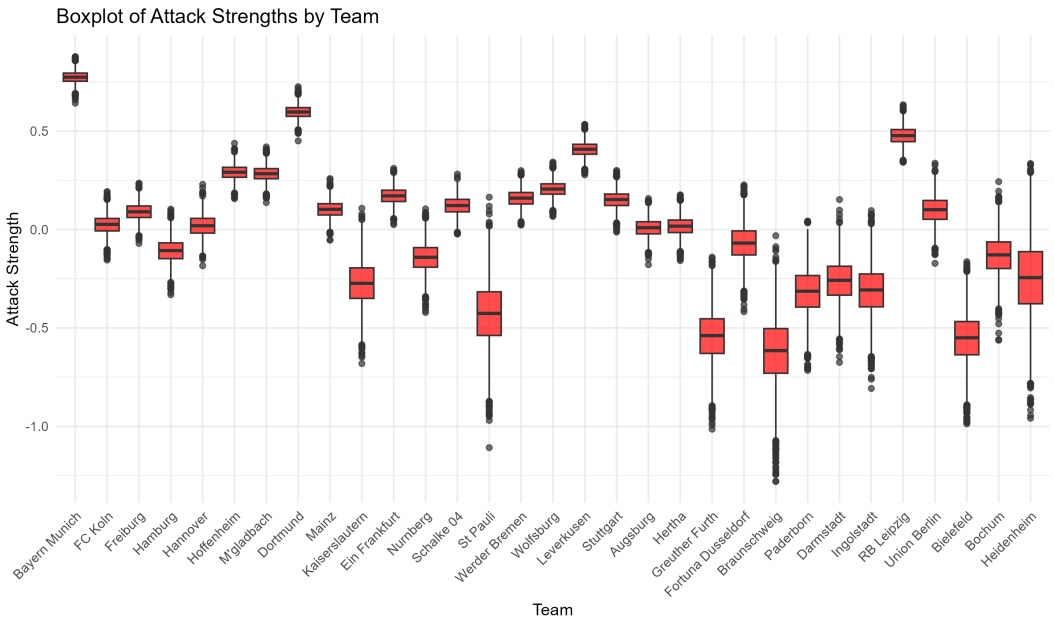

Figure 9: Attack parameters for Baio-Blangiardo model

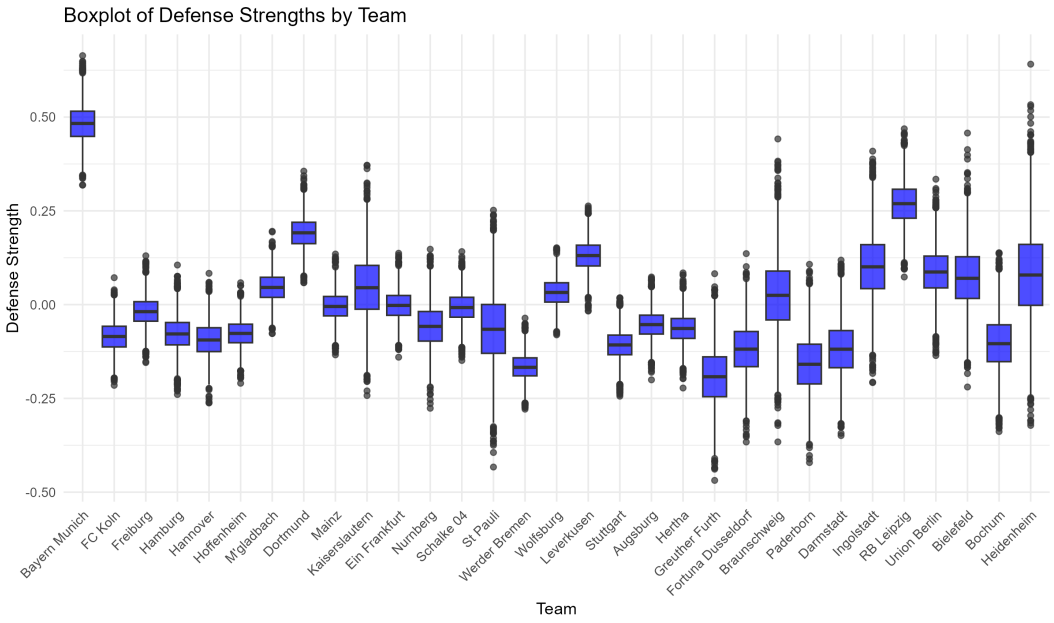

Figure 10: Defense parameters for Baio-Blangiardo model

5.3 Extended Bayesian Model

RPS	Accuracy
0.419	0.491

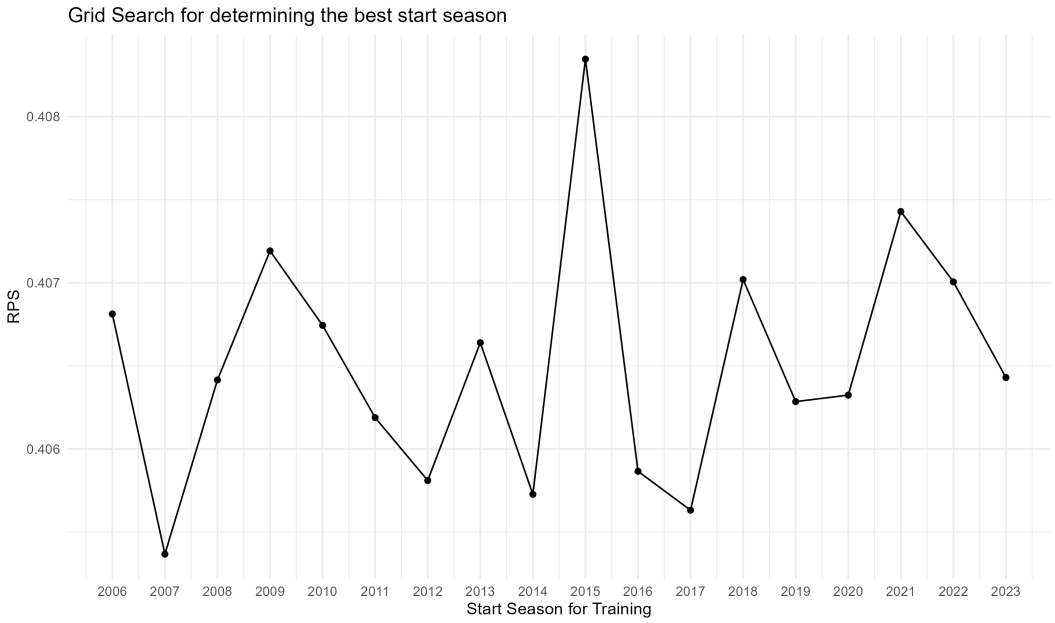

Figure 11: Grid Search for best start season for extended Bayesian model

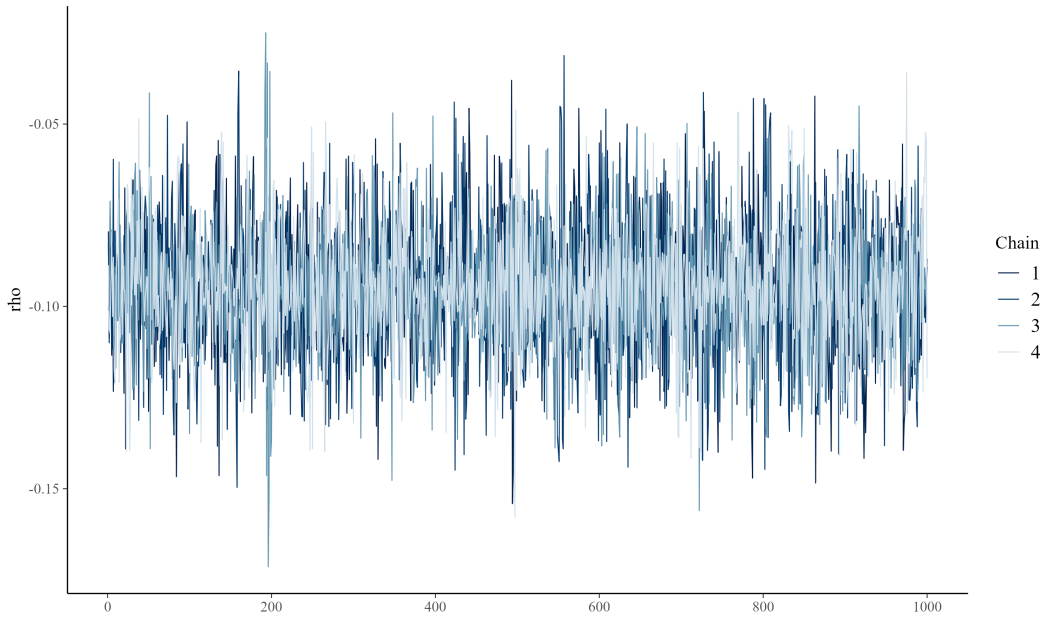

Figure 12: Trace plot for Markov chain Monte Carlo for extended Bayesian model

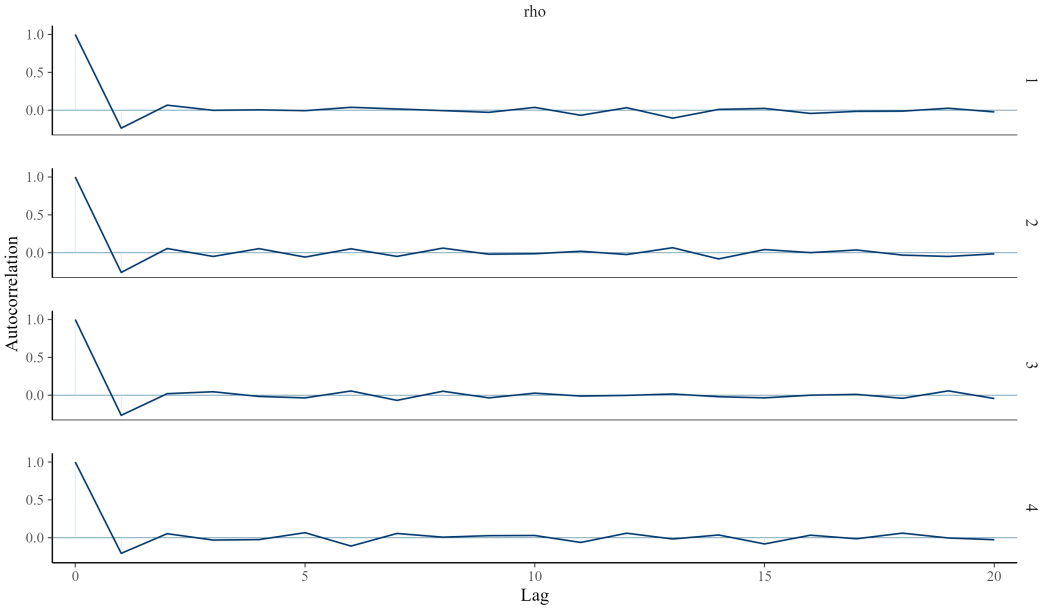

Figure 13: Autocorrelation for Markov chain Monte Carlo for extended Bayesian model

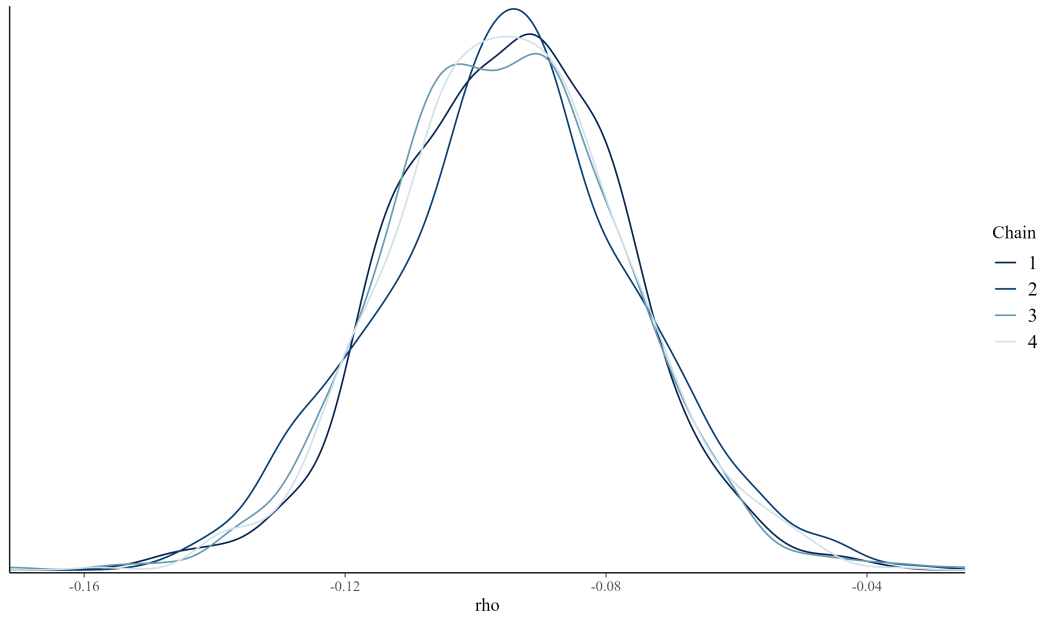

Figure 14: Distribution overlay for Markov chain Monte Carlo for extended Bayesian model

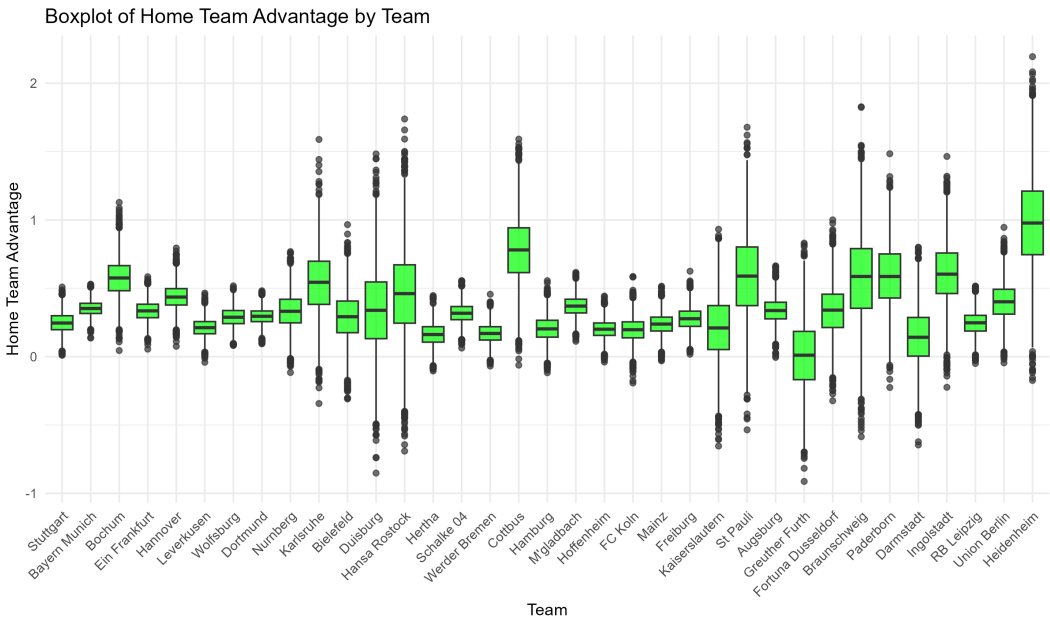

Figure 15: Home team advantage parameters for extended Bayesian model

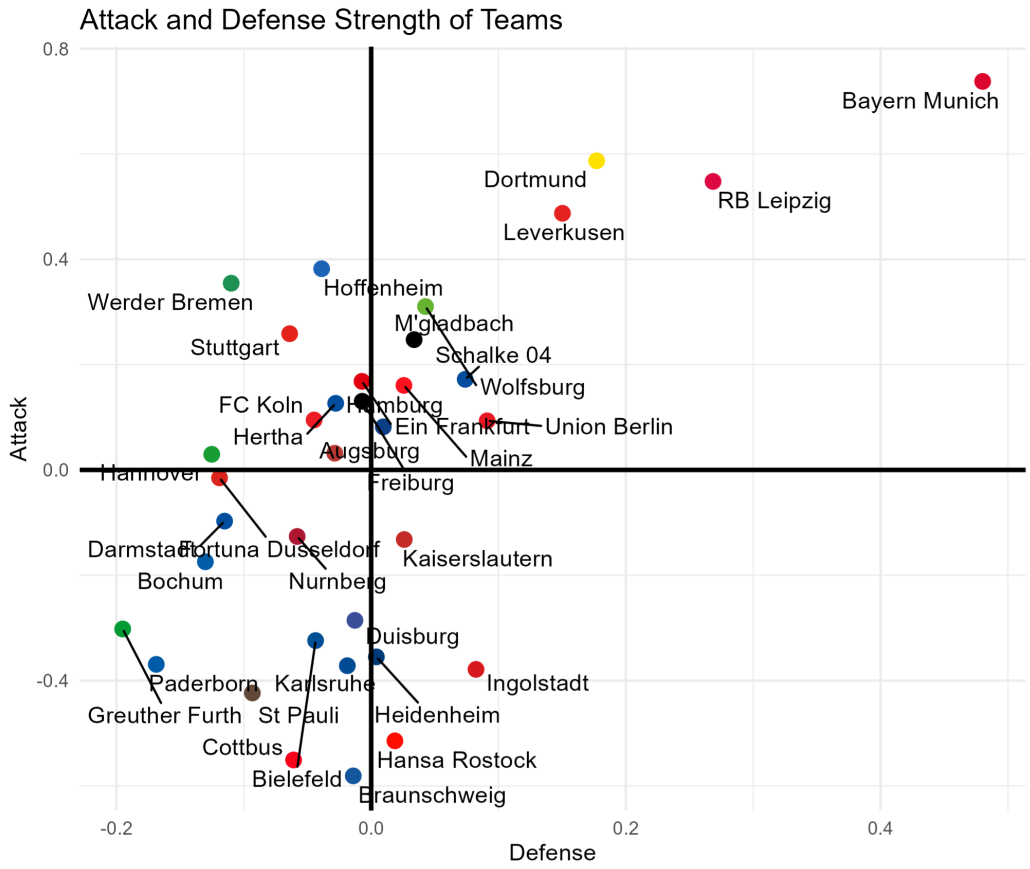

Figure 16: Parameter plot for extended Bayesian model

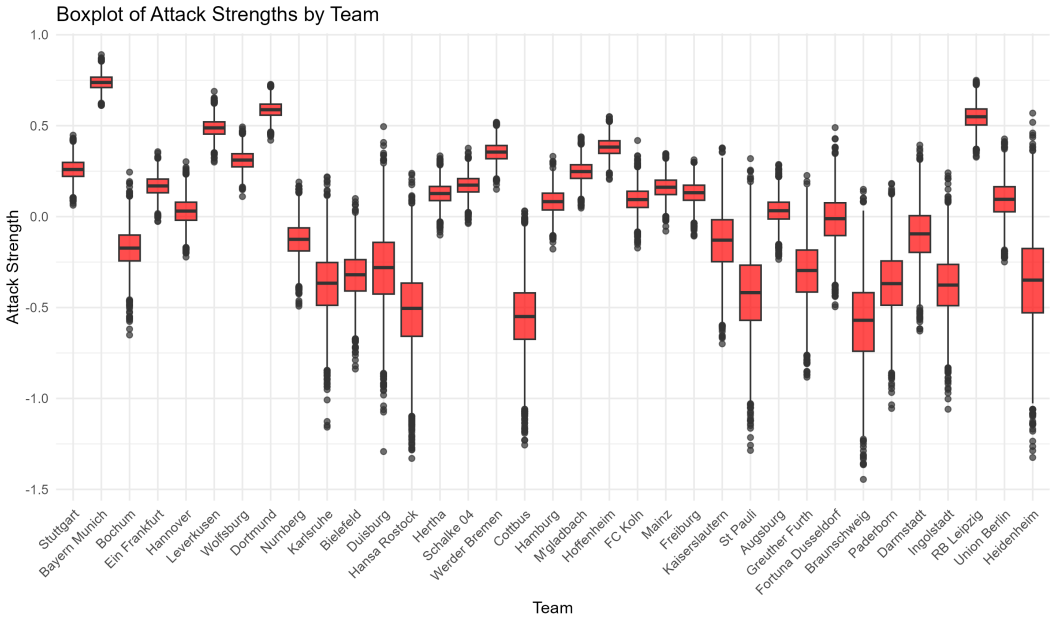

Figure 17: Attack parameters for extended Bayesian model

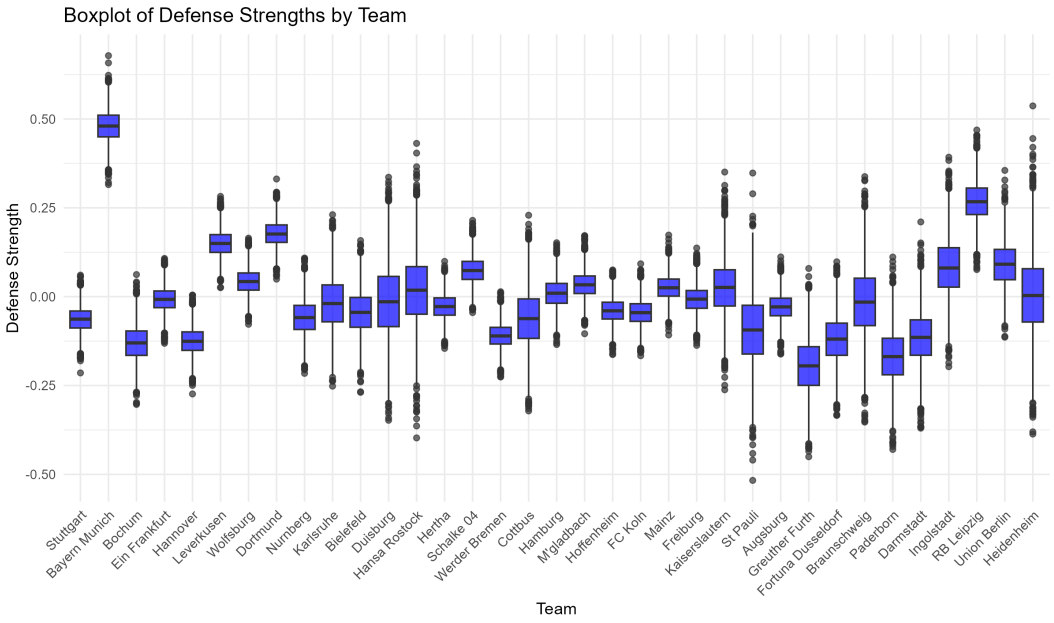

Figure 18: Defense parameters for extended Bayesian model

5.4 Neural Network Model

RPS	Accuracy
0.400	0.515

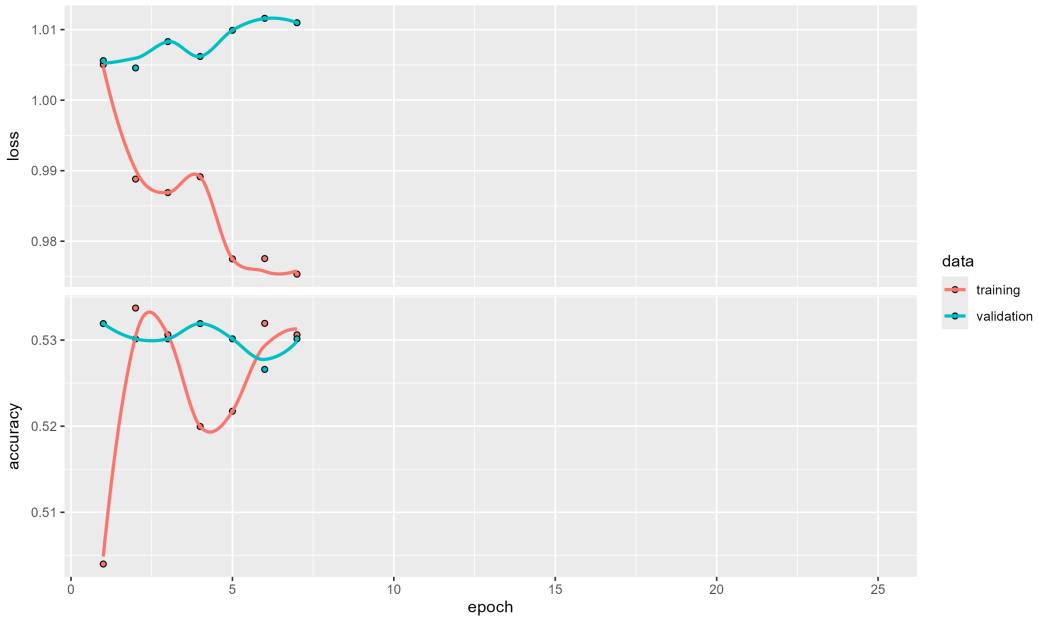

Figure 19: Training of the Neural Network

5.5 Comparison

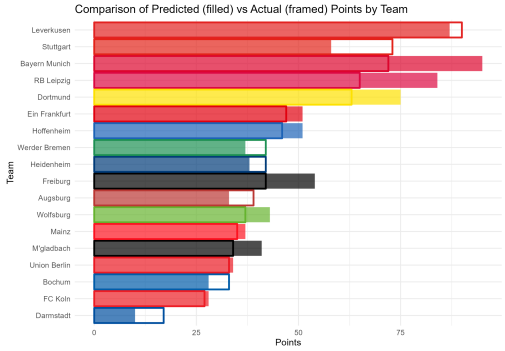

Figure 20: Dixon-Coles Model

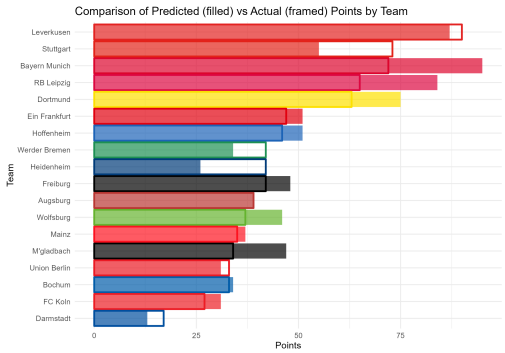

Figure 21: Baio-Blangiardo Model

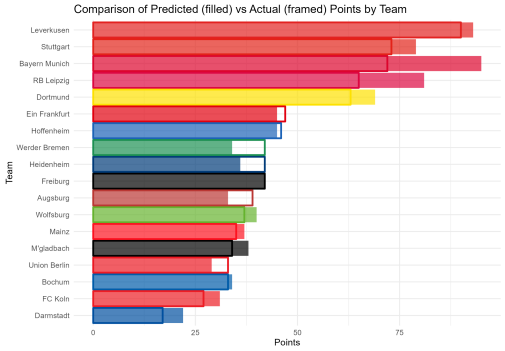

Figure 22: Neural Network Model

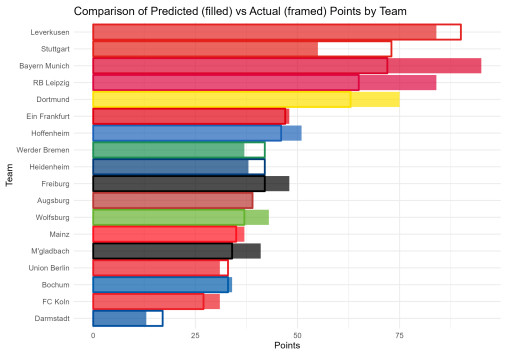

Figure 23: Extended Bayesian Model

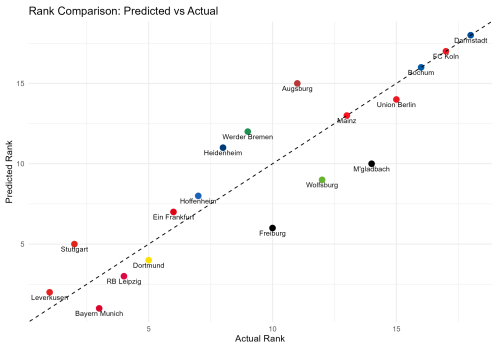

Figure 24: Dixon-Coles Model

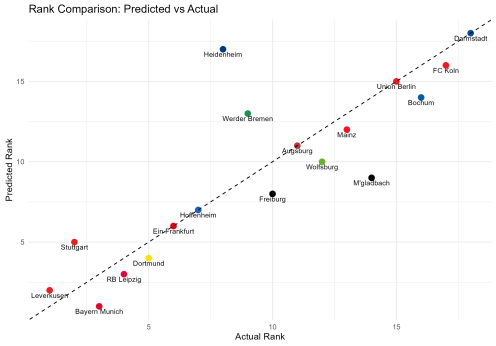

Figure 25: Baio-Blangiardo Model

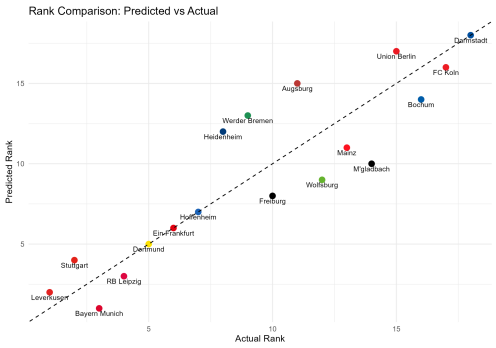

Figure 26: Neural Network Model

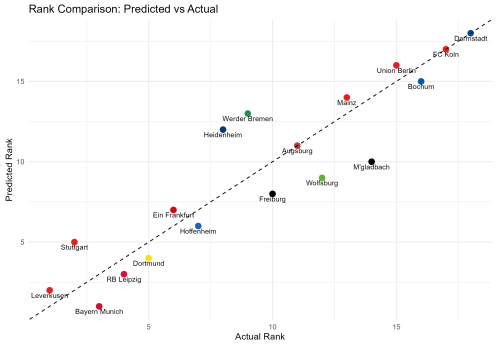

Figure 27: Extended Bayesian Model

Model	RPS	Accuracy	MAE Points	MAE Rank
Dixon-Coles	0.410	0.485	7.611	1.778
Baio-Blangiardo	0.424	0.503	8.278	1.889
Extended Bayesian	0.419	0.491	6.944	1.667
Neural Network	0.400	0.515	5.556	1.889

6 Conclusion