Bayesian Models and Football Forecasting

Booking the Bookie? The Case for the Bayesian Framework

Luis Kaiser, PhD Student in Quant. Finance Matriculation number: 6353366

March 2025

1 Introduction

For finance researchers, sports betting markets offer an interesting micro-environment: They represent a bounded, small-scale financial market where the products are plentiful, liquid, and mostly available to informed buyers who can only go long. (Makropoulou and Markellos 2011) Several papers already demonstrated the inefficiencies on different scales of these markets, most recently highlighting the non-transitivity of bets. (Ours 2024) Trading strategies frequently exploit discrepancies in odds across different bookmakers by simultaneously betting on win, lose, and draw outcomes to secure a safe return. However, strategies based on proprietary prediction models for football games (EV-betting) are less common due to several factors. Football matches are notoriously difficult to forecast, prompting sports betting companies to employ large teams to develop robust models. On top of that, to make profitable bets, trading strategies typically compare market odds with the implied odds of proprietary models. (Altmann 2004) Uncertainty quantification is an integral part when one wants to employ these strategies in practice. Recently, Bayesian methodologies and frameworks have proven to be competitive in this regard. (Ridall, Titman, and Pettitt 2024) This paper argues for using Bayesian models in these betting markets through a twofold rationale: 1) the potential for utilizing a season-based prior, and 2) the ability to fully capture the uncertainty of the pointwise probability estimates of winning and losing. This latter aspect is crucial when constructing betting strategies based on these predictions. The remainder of the paper is organized as follows: First, the dataset construction is described briefly. Next, the paper describes and contrasts the benchmark models with the two Bayesian models. The benchmarks include random forests and frequentist multinomial logit models. The Bayesian approaches comprise a hierarchical ordinal logit model with a season-based prior and a hierarchical ordinal logit model without the season-based prior. In the final section, the results of these methods are discussed, and a basic betting strategy based on credible intervals is introduced and compared against a crude trading strategy based on the random forest model.

2 Dataset

For the task at hand, I have restricted my analysis to the English Premier League from 2018 to the beginning of March this year. The data gathering, cleaning, and feature engineering are carried out in Python. The dataset construction was quite involved from a programming perspective, I omit a full discussion here. The dataset used for training consists of a unified, match-level collection that includes both basic and advanced statistics from 2018 to the beginning of this season. In sum, the training set contains 2340 individual matches. The test set consists of matches from the 2024–2025 season, totaling around 240 games (the exact figure depends on when I last ran my scraper script). For the prediction exercise, I construct features that are known prior to each match, mostly form-based averages per team and the pi-rating (see Hubácek, Sourek, and Zelezny 2019, Yeung et al. 2024). The pi-rating is a measure of team strength that takes home advantage into account. Naturally, performance improves with more sophisticated feature engineering, but this should not be the main focus of this paper. I do, however, place special emphasis on respecting the panel data structure of the dataset and avoiding data leakage. For the Bayesian models in this code, I use a principal component analysis to reduce runtime. The principal components explain around 80% of the variation in the explanatory variables. Introducing more principal components boosts the predictive power of these Bayesian models in my setup but increases runtime drastically. The entire analysis critically depends on the integrity of the datasets and the reliability of the data sources. I will revisit this aspect in the results section. On top of that, I download another dataset on historical closing odds from the major bookmakers in the United Kingdom. This dataset is used to evaluate the performance of each individual algorithm, but crucially, it is not used for prediction. The reason is that, in practice, it is very hard (or very costly) to obtain up-to-date closing odds for upcoming games, and I want this setup to be realistic.

3 Models

The outcome variable in this analysis is categorical, indicating whether a match results in a home win, a draw, or an away win. Therefore, a multinomial logistic regression is the natural starting point and the main focus of this paper.

3.1 The Case for Bayesian Methods

Before delving into the details of the models, it is important to establish the merits for using Bayesian methods in this type of predictive problem. Consider multinomial logistic regression and random forests as baseline models. These models generate probability estimates that can be transformed into implied betting odds, though recalibration may be necessary (see Chen, Gamage, and Ryan 2022). The primary objective is not only to predict the correct outcome class but to generate reliable probability estimates. Moreover, point estimates alone are insufficient. Effective risk management requires assessing the extent to which the model's predictions differ from market odds. For the standard multinomial logit model, uncertainty can be quantified using the inverse of the information matrix in combination with the delta method. However, this approach relies on asymptotic approximations, which may pose challenges when sample sizes are limited. In the case of random forests, uncertainty can be estimated through methods such as analyzing the distribution of class predictions from individual trees and evaluating out-of-bag prediction errors (see Biau and Scornet 2016). While these techniques work reasonably well, they often depend on additional assumptions or require case-specific, ad-hoc adjustments. (Bayesian methods have gained traction in recent years, particularly for enhancing uncertainty quantification in random forests) Given a dataset and specified prior distributions, Bayesian approaches yield a full posterior distribution over parameter values. These posteriors can then be used to derive the posterior predictive distribution of the target variable (see Van de Schoot et al. 2021). Essentially, this extends the benefits of parameter inference uncertainty quantification to the predictive framework without the need for additional assumptions or additional modifications. Uncertainty is inherently captured within the Bayesian model structure. Another key advantage of Bayesian methods is their ability to systematically incorporate prior information. A common challenge in form-based prediction models is that feature values are derived from a team's recent matches, making early-season predictions less reliable due to the influence of the previous season's data. Bayesian methods are particularly well-suited for modeling soccer match outcomes, as they account for both the randomness inherent in individual results and the underlying relative team strength of the competing sides. While capturing relative team strength systematically can be difficult in a frequentist framework, it naturally integrates into the Bayesian paradigm, as will be demonstrated in the following sections.

3.2 Benchmark Models

3.2.1 Random Forests

Random Forests are a widely used method for prediction tasks, having gained considerable popularity because they typically deliver robust performance with minimal hyperparameter tuning. Originally refined by Breiman 2001, the framework builds on classification and regression trees. In its basic form, it trains multiple deep decision trees on bootstrapped subsamples of the dataset (see Hubácek, Sourek, and Zelezny 2019). At each split, a random subset of explanatory variables is considered to reduce correlation among individual trees. A majority vote across all trees then makes the final prediction. Since Random Forests are not the primary focus of this paper, I leave it at this description. Still, in the model construction, hyperparameter tuning is emphasized to make Random Forests a competitive benchmark. I consider a grid of key parameters— maximum tree depth, minimum leaf size, and the number of randomly selected features per split—and identify the best set via 5-fold cross-validation. Special attention is paid to maintaining the temporal structure of the data throughout this procedure by using a time series-specific cross-validation.

3.3 Multinomial Logit

I also considered a multinomial logit model as a second benchmark, as my Bayesian models are essentially multinomial logit models. The multinomial logistic regression model generalizes the binomial logit model by modeling the probability of each class as follows:

$$P(y = k|\mathbf{x}) = \frac{\exp\bigl(\beta_{k0} + \boldsymbol{\beta}_k^{\mathsf{T}}\mathbf{x}\bigr)}{\sum \ell = 1^{K} \exp\bigl(\beta_{\ell 0} + \boldsymbol{\beta}_\ell^{\mathsf{T}}\mathbf{x}\bigr)} \tag{1}$$

where β_{k0} is the intercept for class k, and β_k is the corresponding coefficient vector. For this paper, the outcome classes are win, lose or draw. Each class has its own vector of parameters β, and the (latent) decision boundary is modeled linearly. As the name suggests, the outcome variable follows a multinomial distribution, and estimation is typically carried out using the maximum likelihood method. Standard errors are obtained via the information matrix equality. It is important to note that this model does not automatically select regressors, in contrast to the random forest, for example. In theory, this model would require a variable selection algorithm or a regularization parameter for optimal application as it might be prone to overfitting.

3.4 Hierarchical Ordinal Logit Model (HOLM)

The baseline method of a Bayesian multinomial logit model is the Hierarchical Bayesian Ordered Logistic Regression, similar to what we did in class. For this first model, team strength posteriors are modeled over all seasons together. (for reference, see Rossi 2014)

3.5 HOL model setup

Again, the model's basic structure is equivalent to equation 1. The input variables however are different: For each match, a latent score is calculated in the following way:

$$\eta_n = (\mathrm{s}_{\text{home\_team}_i} + \text{home\_adv}) - \mathrm{s}_{\text{away\_team}_n} + X_n \cdot \beta \tag{2}$$

- s is the latent strength of team i.
- *home_adv* captures home advantage constant across teams
- X_n represents the principal components considered
- β are regression coefficients for principal components.

So, the main difference is the introduction of the team-specific, latent strength parameter and a home advantage parameter, which is constant for all teams and seasons. On top of that, the model has fewer covariates as the principal components reduce the number of covariates drastically. (There are papers that show that for the pandemic seasons, the home advantage vanished; this would be an obvious extension of this setup) The logistic link function maps this latent score into the probabilities domain. As my model has three distinct, ordered outcomes, I need two cutpoints. These cutpoints are also parameters and, therefore, have a prior and posterior distribution.

$$
\begin{aligned}
P(y_n = 1) = P(\eta_n \leq c_1) &= \text{logistic}(c_1 - \eta_n) && (3)\\
P(y_n = 2) = P(c_1 < \eta_n \leq c_2) &= \text{logistic}(c_2 - \eta_n) - \text{logistic}(c_1 - \eta_n) && (4)\\
P(y_n = 3) = P(\eta_n > c_2) &= 1 - \text{logistic}(c_2 - \eta_n) && (5)
\end{aligned}
$$

with

$$\text{logistic}(x) = \frac{1}{1 + e^{-x}} \tag{6}$$

The full conditional likelihood function is then obtained as:

$$
\begin{aligned}
L(\theta) &= \prod_{n=1}^{N} P(y_n \mid \eta_n, c) && (7)\\
&= \prod_{n=1}^{N} \begin{cases} \text{logistic}(c_1 - \eta_n), & \text{if } y_n = 1 \text{ (Away Win)} \\ \text{logistic}(c_2 - \eta_n) - \text{logistic}(c_1 - \eta_n), & \text{if } y_n = 2 \text{ (Draw)} \\ 1 - \text{logistic}(c_2 - \eta_n), & \text{if } y_n = 3 \text{ (Home Win)} \end{cases} && (8)
\end{aligned}
$$

As the posterior has no closed form, I am using MCMC algorithm to assess posteriors numerically. The code uses Stan's default sampler, which is the No-U-Turn Sampler (NUTS)—a variant of Hamiltonian Monte Carlo (HMC). For this baseline model, convergence and mixing characteristics seem to be already fine for shorter chains around 1000 iterations, but in the end I decided on 5000 iterations with 1000 burn-in iterations.

3.5.1 Priors of HOLM

Each team i has a latent team-strength parameter s_i, which is modeled with the following prior.

$$s_i \sim N(0, \sigma_i) \tag{9}$$

where σ_i is the hyperparameter modeled as:

$$\sigma_i \sim N^{+}(0, 5) \tag{10}$$

The first layer of the hierarchy is the variance of the strength parameter, which differs then for each teams prior. This accounts for the difference in teams consistency with their individual strength. The cutpoints prior is modeled as $N \sim (0, 5)$ and the priors on the regressors are modeled as $\sim N(0, 4)$. During the fine-tuning of the models, I adjusted the prior variance at some point. This description here contained the baseline. The goal of this setup is to introduce fairly uninformative priors. One problem with the initial setup is the number of features. With well over 50 partially highly correlated features per match, the mixing of chains is poor, computation time is high, and chains need to be quite in order to ensure identification. For that reason, I decided on using a principal component analysis on the features in order to speed up computation time. In the end, I stuck with nine principal components which account for around 80 percent of the variation.

3.5.2 Convergence, Mixing and Posteriors of HOLM

Already, for smaller iteration sizes and burn-in numbers, mixing and convergence seem to be fine. The following trace plots show clear mixing and convergence, with a high variance in posterior distributions. One can also see in table 1 that the effective sample size is large and the Rhat parameter is 1, indicating converged chains.

Table 1: Summary MCMC Estimation

	Parameter	Mean	SE Mean	Effective Sample Size (N_eff)	Rhat
1	home_adv	0.127	0.032	7,043.953	1.000
2	sigma_team	0.557	0.001	5,738.059	1.000
3	c[1]	-0.657	0.032	7,040.891	1.000
4	c[2]	0.391	0.032	7,051.416	1.000
5	beta[1]	0.005	0.0001	14,823.570	1.000

One aspect I want to highlight is the overlapping and close posteriors of the cutpoints, as showcased in figure 2. This hints on the fact that the probability of games ending in a draw is small and hard to identify, which might be a distinguishing factor of model performance. Other posteriors are also produced in the code, but omited in this paper for brevity. The performance of this iteration is already on par with frequentist benchmarks while offering a better uncertainty quantification and better probability estimates especially for draws. (See section 5) In the next section, I want to leverage more the structure of the Bayesian framework by including season-based priors.

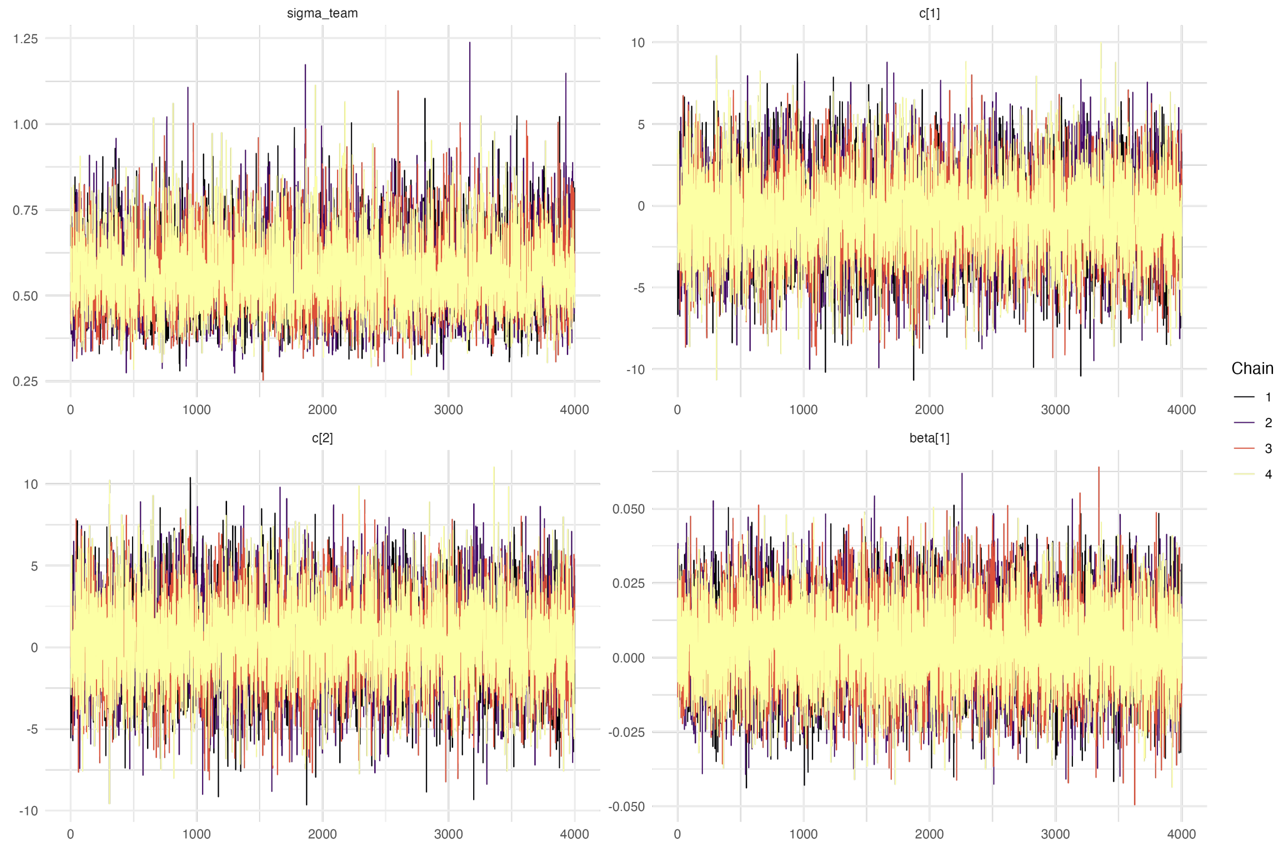

Figure 1: MCMC Trace Plots for Key Parameters

3.6 HOLM with Season Based Prior

HOL models lack a proper solution to incorporate information at the beginning of the season. I introduce season-based priors to solve that and showcase the flexibility of the Bayesian framework. The prior of each team in the first season is still modeled as:

$$s_{i,1} \sim N(0, \sigma_i) \tag{11}$$

$$\sigma_i \sim N^+(0, 5) \tag{12}$$

For each season that follows, the mean of the team strength hyperparameter is adjusted accordingly:

$$s_{i,t+1} \sim N(\mu_{i,t}, \sigma_{i,t}) \tag{13}$$

with $\mu_{i,t}$ being derived from the final standings of the period beforehand, standardized to lie between 0 and 1. Besides this change, this model is virtually unchanged to what I described earlier. This implementation introduces a season-based posterior for team strength, which I expected to increase performance while also increasing computational burden, as it introduces 20 new posteriors per season to simulate. In sum, this introduces around 120 additional parameters to estimate. I acknowledge the similarity of including a ”fudge” factor at the season’s beginning in a frequentist model.

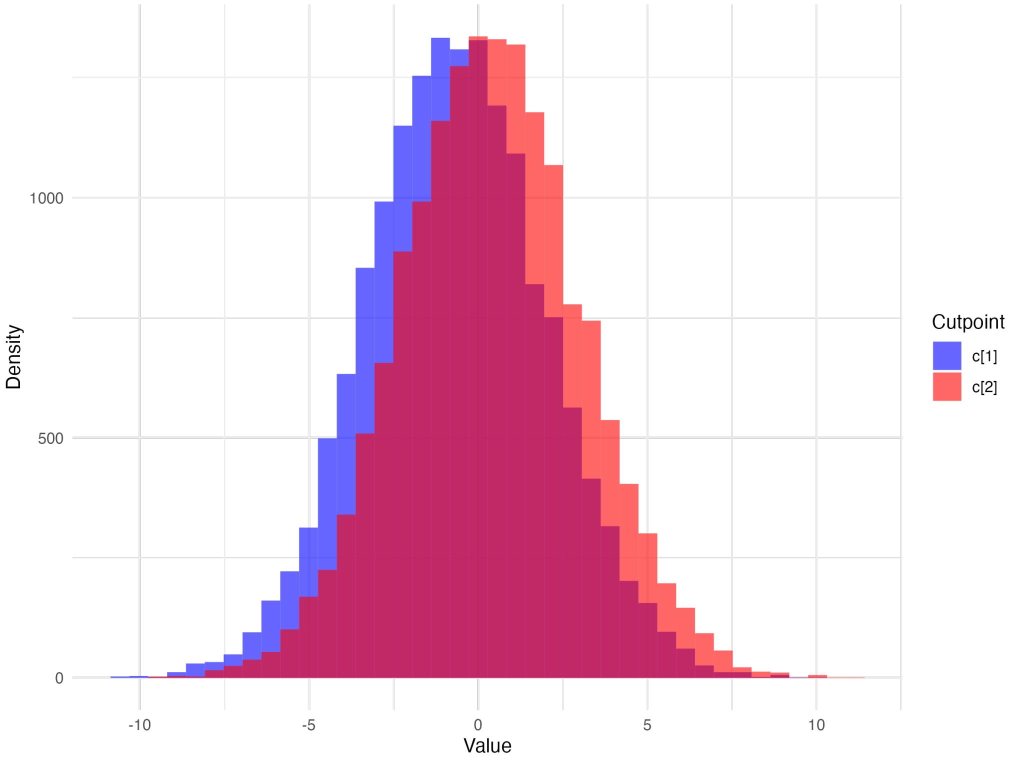

Figure 2: Overlapping Posterior Distributions of Cutpoints $c[1]$ and $c[2]$

3.6.1 Convergence, Mixing and Posteriors of HOLM with Season Based Priors

As expected, season-based priors introduce a higher dimensionality, and thus, smaller chains are not converging anymore. However, using chains with 5000 iterations and 1000 burn-in iterations is still enough. Again, the convergence variables indicate convergence, and the trace plots look fine (see figure 4). The season-based prior should also allow us to outperform the competitive benchmark, especially at the beginning of the season. One can see in figure 3 an example of the team strength posterior of an example team (Manchester United) over the seasons. This posterior plot is very reassuring, when considering the historical performance of Manchester

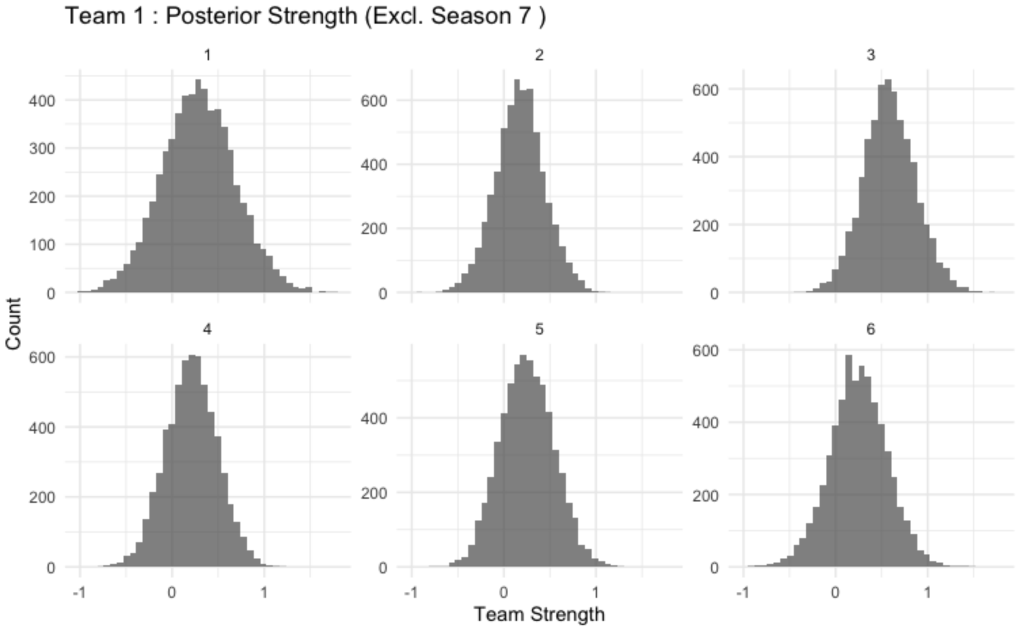

Figure 3: Posterior distribution of team strength of Arsenal

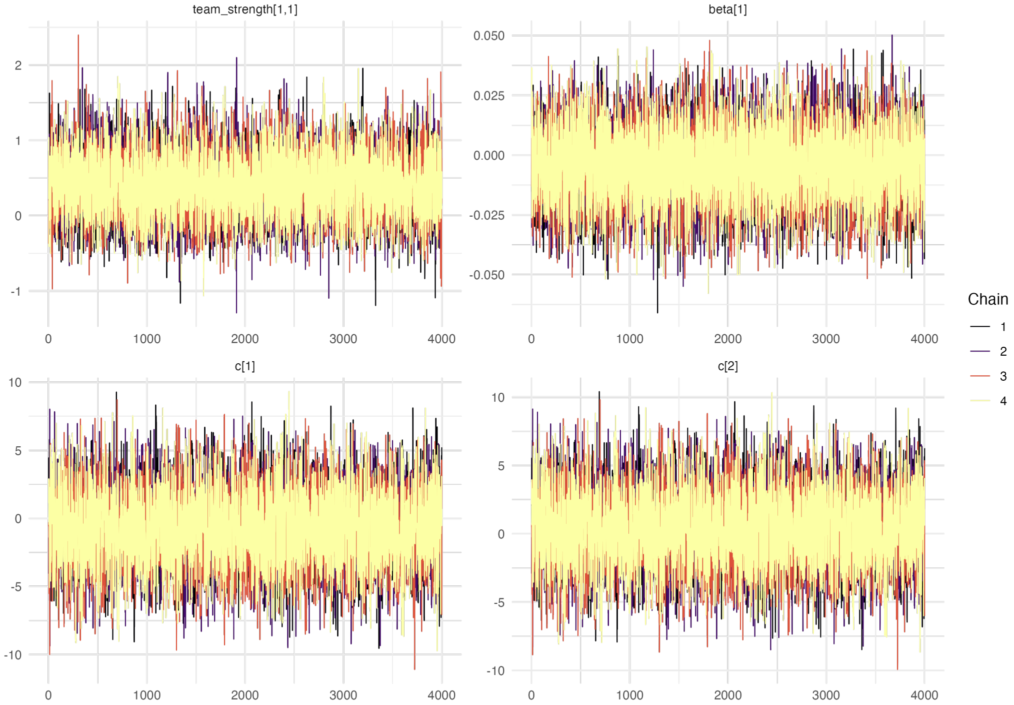

Figure 4: MCMC Trace Plots for Key Parameters

United over the last few seasons. As a robustness checks, posterior means for the this season across all teams are plotted in the script for this model. The best performing teams at the top, the worse performing teams at the bottom, which serves as a sanity check. It seems that modeling team strength season by season has merit, as there are considerable differences in means and support of the team strength for each season. To flip the argument: Modeling the team strength posterior over the whole dataset might introduce too much regularization.

4 Results

4.1 Performance comparison

On the test dataset, the HOLM model with season-based priors performs best with a prediction accuracy of 49%(and also the best F-Score), while the Random forest algorithm has a prediction accuracy of 44%, the logit model of around 48% and the baseline HOLM model of 40%. In sample, the HOLM model with season based priors also performs best. Notably, all models fail to forecast draws at all, which was expected, as draws are never the popular class. Simple accuracy of predictions is not a good way to evaluate the performance of the measures. The F1 score combines false positive, false negative and prediction accuracy into one metric, but it also fails to assess what the goal of this prediction exercise is.

The goal is to predict the probabilities with high accuracy, not to always have the right prediction. Building a performance metric for that goal is not straightforward. I construct a mean implied bookmaker probability measure from the odds dataset, which is not part of the test or training data, in order to judge how well each individual forecasting algorithm performs in comparison to the mean bookmaker implied probability. The implied probabilities of the closing odds are inherently biased, as they contain the bookmakers margin. There are many different possibilities to debias the closing odds (Hubácek, Sourek, and Zelezny 2019); I chose the most simple one, which just ensures that the implied probabilities add up to one for every individual bookmaker. (In practice, bookmaker do have a systematic margin, i.e., if one bets on all possible outcomes at the same time with one bookmaker, he loses money, even before considering taxes.) The mean implied probability over the different betting places serves as a benchmark. A "good" model should produce probabilities that are not too far off the implied mean probabilities in general. I compute the mean absolute difference of the point prediction of probabilities of each outcome for every match in the test set and the debiased implied mean probabilities As can be seen in figure 5, HOL models improve over the logit benchmarks, especially in better accuracy for draw prediction while still being on par with the benchmark models for prediction accuracy for home and away win probabilities. The baseline HOLM (Base Bayes) is outperformed by the random forest for home-win and away-win predictions. The HOL model with season-based priors (”Bayes Adv”) performs on par (slightly worse) with the random forest for home and away team win probability prediction but clearly outperforms for draw probability predictions. When assessing the performance metric over game weeks in the test set, it seems that the Bayesian models do not have as much variance in performance compared to the benchmarks. (Plot is produced within the scripts)

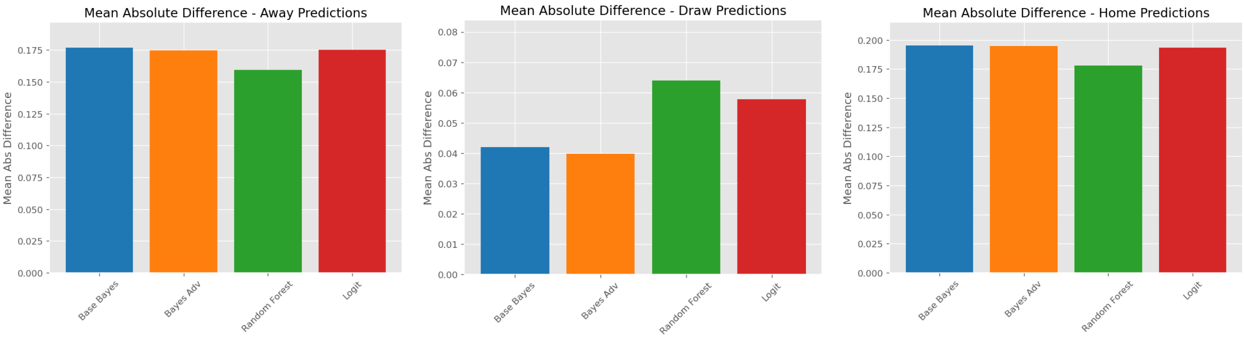

Figure 5: Absolute Difference Implied Probabilties and Models

4.2 Credible Intervals and trading strategies

The last part of this paper should give a brief look at why and how this all matters in practice. One important advantage of Bayesian models is systematic uncertainty quantification. Figure 6 showcases the 95% credible intervals together with the bookmakers implied probability for a random subset of games. Interestingly, the ”draw” credible intervals are bounded by one-third, as a draw is never the most probable outcome in the models. Notably, there are cases where the implied bookmaker’s probabilities lie outside of that prediction of credible interviews. To showcase the advantage credible intervals offer in the last part of this, I want to showcase a simplistic way of using this observation to build a betting strategy. **Please especially note section A on the caveats of this approach, this is not a recommendation of actually using money on this type of strategy.**

This betting strategy aims to exploit odds, which are exceedingly high. Exceedingly high odds mean implied probabilities, which are extraordinarily low. To identify exceedingly low probabilities, the strategy compares the lower bound of the HOLM with season-based priors credible intervals with the implied probability one particular bookmaker. After identifying exceedingly high odds, the model simply goes long with this specific outcome. As one cannot go short in betting markets, one cannot exploit exceedingly low odds. I backtested dynamic staking (invest 2% of wealth for each bet) and fixed staking (0.5 units) for this strategy and the benchmark strategy. For the benchmark strategy I implemented a similar strategy to the random forest, but using a fixed percent difference (25%) between the random forest probability forecast and the implied bookmakers probability. The results are displayed in table 2. While the random forest betting strategy will place more or less bets, depending on the percentage differences imposed, the credible intervalls dynamically change the distance ”needed”, depending on the individual match. In this simplistic backtest, one can observe (see figure 7) that both the credible-interval based as well as the random forest-based strategy can be profitable. Dynamic staking outperforms in this setup, even though the staking is not optimally carried out. (See Thorp 2011 for a discussion of optimal staking) To end this section I want to again reference section A for a discussion of why these results might be not generalizable easily. Results also might change with different staking strategies, credible interval percentiles and distances for the random forest.

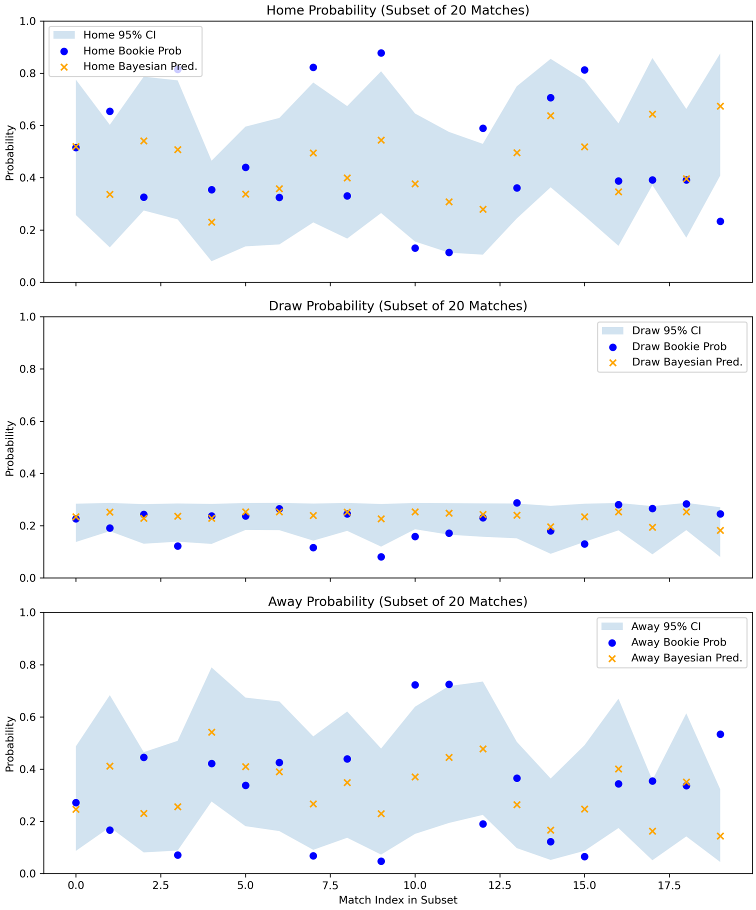

Figure 6: Credible Intervalls of 20 random matches in the test dataset

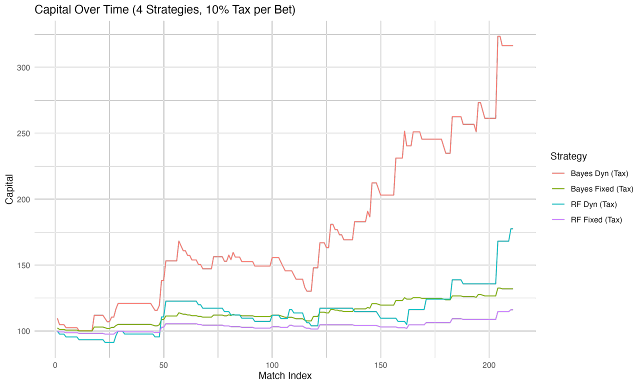

Figure 7: Capital over time with dynamic betting of 5 percent of the wealth

Table 2: Comparison of 4 Betting Strategies with 10 percent Bet Tax

Strategy	Num_Bets	Wins	Losses	Avg_Odds	Final_Capital	Profit	ROI
Bayesian (Dynamic, w. Tax)	72	25	47	6.944	316.546	216.546	2.165
RF (Dynamic, w. Tax)	36	12	24	7.549	177.694	77.694	0.777
Bayesian (Fixed, w. Tax)	72	25	47	6.944	132.145	32.145	0.321
RF (Fixed, w. Tax)	36	12	24	7.549	116.265	16.265	0.163

5 Conclusion

Forecasting soccer games offers a natural use case for Bayesian methods which I showcased in this paper. If one wants to bet profitably in the long term, an assessment of ”how far off” the own model might be from the implied probabilities of the bookies is essential. This paper showed how fairly simplistic Bayesian models can accomplish that, while still being competitive in terms of performance. The full posterior for the outcome probabilities is a key advantage of the Bayesian framework which is the main contribution to results. While the baseline Bayesian model offers an improvement in terms of performance over the baseline multinomial logit, it falls short against a well tuned random forest algorithm. The HOL model with season based priors improves in terms of prediction accuracy over all models and offers slightly better performance in terms of distance to benchmark probabilities. While the trading strategy backtest has to be considered with caution it showcases the potential of using credible interval based strategies for trading.

References

Altmann, A. (Jan. 2004). “A statistical approach to sports betting”. Unpublished. URL: https://openaccess.city.ac.uk/id/eprint/8431/.

Biau, Gérard and Erwan Scornet (2016). “A random forest guided tour”. In: *Test* 25.2, pp. 197–227.

Breiman, Leo (2001). “Random forests”. In: *Machine learning* 45, pp. 5–32.

Chen, Lihua, Prabhashi Withana Gamage, and John Ryan (2022). “Debias random forest regression predictors”. In: *Journal of statistical research* 56.2, pp. 115–131.

Hubácek, Ondrej, G Sourek, and F Zelezny (2019). “Score-based soccer match outcome modeling–an experimental review”. In: *MathSport International.*

Makropoulou, Vasiliki and Raphael N Markellos (2011). “Optimal price setting in fixed-odds betting markets under information uncertainty”. In: *Scottish Journal of Political Economy* 58.4, pp. 519–536.

Ours, Jan C van (2024). “Nontransitive Patterns in Long-Term Football Rivalries”. In: *Journal of Sports Economics* 25.7, pp. 802–826.

Ridall, P Gareth, Andrew C Titman, and Anthony N Pettitt (Dec. 2024). “Bayesian state-space models for the modelling and prediction of the results of English Premier League football”. In: *Journal of the Royal Statistical Society Series C: Applied Statistics*, qlae075. ISSN: 0035-9254. DOI: 10.1093/jrsssc/qlae075. eprint: https://academic.oup.com/jrsssc/advance-article-pdf/doi/10.1093/jrsssc/qlae075/61250598/qlae075.pdf. URL: https://doi.org/10.1093/jrsssc/qlae075.

Rossi, Peter (2014). *Bayesian non-and semi-parametric methods and applications.* Princeton University Press.

Thorp, Edward O (2011). “Understanding the Kelly criterion”. In: *The Kelly capital growth investment criterion: theory and practice.* World Scientific, pp. 509–523.

Van de Schoot, Rens et al. (2021). “Bayesian statistics and modelling”. In: *Nature Reviews Methods Primers* 1.1, p. 1.

Yeung, Calvin et al. (2024). “Evaluating soccer match prediction models: a deep learning approach and feature optimization for gradient-boosted trees”. In: *Machine Learning* 113.10, pp. 7541–7564.

A Caveats on the betting strategy

One should take the results on the betting strategy with a grain of salt. I want to give a non-exhaustive list of limitations and restrictions of these betting strategies:

- The betting strategies are a form of EV-betting. These are not abitrage-betting strategies and only ”guarantee” to profit in the long run, if carried out properly. Only because on this specific subset, a profit is calculated in the backtest does not mean that in the future these strategies would be profitable.
- The odds dataset consists of closing odds (the odds directly before the game). These are typically the type of odds with the most mispricing as football fans typically do bet directly before the game in a skewed manner towards popular teams.
- The backtest relies on the assumption that the odds used in this project are accurate and readily available.
- Taxes are only incorporated in a crude way and may differ from country to country.

Nonetheless, this showcases the potential of credible intervals: They offer a reliable way to judge the distance between the bookmaker's implied probabilities and the model probabilities, which can be exploited with trading strategies.