FACULDADE
DE ECONOMIA
UNIVERSIDADE
DO PORTO

Post-Hoc Calibration Methods for Machine-Learning Forecasts of Football Outcomes

Pedro Manuel Gonçalves Roque

Dissertation

Master in Modelling, Data Analysis and Decision Support Systems

Supervised by:

PhD Jorge Valente

2026

Acknowledgements

First, I would like to express my gratitude to my advisor, Professor Jorge Valente, who, in the face of every challenge, always showed patience, humanity, and kindness - while demonstrating key knowledge in the areas essential to this dissertation.

To my girlfriend Bea, to my brother João, for the abundant help. Thank you.

To my parents, for their knowledge of life's trickery, and constant unconditional love and support, for always reminding me that, no matter what, I will always have soup at their table. Love you.

Abstract

Machine learning classifiers do not generally produce probability estimates that can be interpreted as probabilities, and post-hoc calibration is the standard remedy. Football 1X2 forecasting normally poses a specific difficulty for that remedy - the forecast is a three-class probability vector, whereas the calibration methods most commonly applied in the sports forecasting literature are binary techniques requiring renormalisation. This dissertation compares four post-hoc calibration methods - Platt scaling and isotonic regression applied one-versus-rest, and Dirichlet calibration in both its full and diagonally constrained (vector scaling) forms - applied to Random Forest and XGBoost forecasts across six major European leagues between 2000 and 2025, under strict temporal validation with de-vigged bookmaker odds as the benchmark.

Calibration improves probabilistic performance reliably but modestly, reducing out-of-sample logarithmic score by approximately 0.3%. The choice of method matters an order of magnitude more than the decision to calibrate: the best treatments improve on the uncalibrated forecast by 0.0027 in log loss, while non-parametric isotonic regression degrades it by between 0.0135 and 0.0302 depending on the convention used to floor probabilities before taking logarithms, leaving the calibrated forecast worse than a single-covariate Elo baseline.

The distinction that proves decisive is not the one the literature emphasises. Among the parametric maps the ordering is stable but the separation is negligible: vector scaling attains the lowest logarithmic score for all three base learners, ahead of Platt scaling by 0.0002 in each case and ahead of full Dirichlet calibration by between 0.0002 and 0.0008, and only the last of these excludes zero under a block bootstrap, for one base learner. What separates methods materially is whether the map is parametric at all. A learning-curve analysis finds that at 250 calibration matches every method is worse than no calibration and the parametric maps are ranked by parameter count, separated by 0.0115; that they become beneficial between 500 and 2,000 matches depending on their complexity; and that isotonic regression remains above the uncalibrated benchmark at every calibration-set size available. The ordering by complexity is therefore real throughout, and immaterial once the calibration set is large enough to support the maps being compared. Because the calibration window in a temporally ordered pipeline is bounded by the tension between sample size and representativeness, this constraint is intrinsic

to the forecasting problem rather than incidental to the design.

Expected Calibration Error ranks forecasters in systematic contradiction to every strictly proper scoring rule examined, because it rewards forecasts that are reliable but uninformative. And class-wise analysis reverses the received view of the draw, as its probabilities are the best calibrated of the three outcomes, because they occupy a narrow band around the base rate, indicating that the widely reported difficulty of the draw class is a deficiency of sharpness rather than of calibration and is therefore invisible to calibration diagnostics. The de-vigged market forecast outperforms every model-calibration combination on every proper scoring rule; calibration narrows that gap by an order of magnitude less than its width, locating the models' deficiency in the information available to them rather than in the fidelity with which they report it.

Keywords: Probability Calibration; Football Forecasting; Dirichlet Calibration; Vector Scaling; Isotonic Regression; Platt Scaling; Proper Scoring Rules; Expected Calibration Error; Sharpness; Machine Learning; XGBoost; Random Forest; Elo Rating; Bookmaker Odds

Resumo

Os classificadores de machine learning não produzem, em geral, estimativas de probabilidade que possam ser interpretadas como probabilidades, sendo a calibração post-hoc o remédio padrão. A previsão do resultado 1X2 no futebol coloca normalmente uma dificuldade específica a esse remédio - a previsão é um vetor de probabilidades de três classes, enquanto os métodos de calibração mais comummente aplicados na literatura de previsão desportiva são técnicas binárias que exigem renormalização. Esta dissertação compara quatro métodos de calibração post-hoc - Platt scaling e regressão isotónica aplicados one-versus-rest, e calibração de Dirichlet nas suas formas completa e diagonalmente restringida (vector scaling) - aplicados a previsões de Random Forest e XGBoost em seis das principais ligas europeias entre 2000 e 2025, sob validação temporal estrita, tendo como referência (benchmark) as odds das casas de apostas ajustadas para remover a margem (de-vigged).

A calibração melhora o desempenho probabilístico de forma consistente, mas modesta, reduzindo o logarithmic score fora da amostra (out-of-sample) em aproximadamente 0,3%. A escolha do método importa uma ordem de grandeza mais do que a decisão de calibrar: os melhores tratamentos melhoram a previsão não calibrada em 0,0027 em log loss, enquanto a regressão isotónica não paramétrica degrada essa previsão entre 0,0135 e 0,0302, consoante a convenção utilizada para limitar inferiormente (floor) as probabilidades antes de se tomar o logaritmo, deixando a previsão calibrada pior do que um baseline Elo de uma única covariável.

A distinção que se revela decisiva não é aquela que a literatura enfatiza. Entre os mapas paramétricos, a ordenação é estável, mas a separação é negligenciável: o vector scaling atinge o logarithmic score mais baixo para os três modelos base, à frente do Platt scaling em 0,0002 em cada caso, e à frente da calibração de Dirichlet completa entre 0,0002 e 0,0008, sendo que apenas este último exclui o zero sob um block bootstrap, e apenas para um dos modelos base. O que separa os métodos de forma material é se o mapa é ou não paramétrico. Uma análise de curva de aprendizagem (learning curve) verifica que, com 250 jogos de calibração, todos os métodos são piores do que a ausência de calibração e os mapas paramétricos ordenam-se pelo número de parâmetros, separados por 0,0115; que se tornam benéficos entre 500 e 2000 jogos, consoante a sua complexidade; e que a regressão isotónica permanece acima do benchmark não calibrado em todas as dimensões de conjunto de calibração disponíveis. A ordenação

por complexidade é, portanto, real ao longo de todo o intervalo - e irrelevante a partir do momento em que o conjunto de calibração é suficientemente grande para suportar os mapas em comparação. Uma vez que a janela de calibração, num pipeline ordenado temporalmente, é limitada pela tensão entre dimensão da amostra e representatividade, esta restrição é intrínseca ao problema de previsão - e não incidental ao desenho do estudo.

O Expected Calibration Error ordena os previsores em contradição sistemática com todas as regras de pontuação estritamente próprias (strictly proper scoring rules) examinadas, porque recompensa previsões que são fiáveis (reliable), mas pouco informativas (uninformative). E a análise por classe inverte a visão estabelecida sobre o empate, uma vez que as suas probabilidades são as melhor calibradas das três classes de resultado, por ocuparem uma faixa estreita em torno da taxa base (base rate), indicando que a dificuldade amplamente reportada da classe do empate é uma deficiência de nitidez (sharpness), e não de calibração, sendo por isso invisível aos diagnósticos de calibração. A previsão do mercado ajustada para remover a margem (de-vigged) supera todas as combinações modelo-calibração em todas as regras de pontuação próprias; a calibração reduz essa diferença numa ordem de grandeza inferior à sua amplitude, situando a deficiência dos modelos na informação de que dispõem - e não na fidelidade com que a reportam.

Palavras-chave: Calibração de Probabilidades; Previsão de Futebol; Calibração de Dirichlet; Vector Scaling; Regressão Isotónica; Platt Scaling; Regras de Pontuação Próprias; Expected Calibration Error; Sharpness; Machine Learning; XGBoost; Random Forest; Elo Rating; Odds de Casas de Apostas

Contents

Acknowledgements . . . iii
Abstract . . . v
Resumo . . . viii

1 Introduction **1**
1.1 Motivation and context . . . 1
1.2 Problem statement . . . 2
1.3 Objectives and research questions . . . 3
1.4 Structure of the dissertation . . . 4

2 Literature Review **5**
2.1 Football forecasting as a probabilistic problem and classical statistical baselines 5
2.1.1 Classical baseline: goal-based Poisson models . . . 5
2.1.2 Alternatives that relax Poisson assumptions . . . 6
2.1.3 Baselines beyond goal models: heuristics, match stats, and crowd signals . . . 6
2.2 Elo and rating systems as team-strength signals . . . 7
2.3 Machine learning for football probability estimation . . . 8
2.4 Proper scoring rules and probabilistic evaluation and calibration . . . 10
2.5 Post-hoc probability calibration . . . 11
2.5.1 Platt scaling . . . 12
2.5.2 Isotonic regression . . . 13
2.5.3 Extending binary calibration to three classes . . . 14
2.5.4 Natively multiclass calibration . . . 14
2.5.5 Assessing calibration . . . 15
2.5.6 Calibration under distribution shift . . . 16
2.5.7 Calibration in football forecasting . . . 17
2.6 Betting odds as probability forecasts, de-vigging and implied probabilities . . 17
2.7 Probability quality as the decision-relevant criterion . . . 18
2.8 On temporal validation, data leakage avoidance, and non-stationarity . . . 19

3 Methodology **21**
3.1 Data . . . 21
3.1.1 Source . . . 21
3.1.2 Competition scope . . . 21
3.1.3 Cleaning . . . 23
3.2 Chronological partitioning . . . 24
3.3 Leakage control . . . 25
3.4 Feature Engineering . . . 26
3.4.1 Construction order . . . 26
3.4.2 Feature set . . . 26
3.4.3 Excluded features . . . 28
3.5 Base models . . . 29
3.5.1 Baselines . . . 29
3.5.2 Machine-learning models and hyperparameter tuning . . . 29
3.6 Post-hoc calibration . . . 33
3.7 Market baseline . . . 34
3.8 Evaluation . . . 35
3.8.1 Proper scoring rules . . . 35
3.8.2 Calibration diagnostics . . . 36
3.8.3 Statistical inference . . . 36
3.9 Robustness analysis . . . 37
3.10 Software and reproducibility . . . 37

4 Results **38**
4.1 Overview . . . 38
4.2 The modelling sample and the market baseline . . . 39
4.2.1 Sample construction . . . 39
4.2.2 Outcome frequencies . . . 39
4.2.3 Shot-based features and their coverage . . . 40
4.2.4 The bookmaker margin . . . 40
4.2.5 Calibration of the market forecast . . . 41
4.3 Uncalibrated base-model performance . . . 42
4.4 The effect of calibration . . . 44
4.4.1 Experimental design . . . 44
4.4.2 Calibration produces small but statistically reliable gains . . . 46
4.4.3 The parametric maps are consistently ordered but negligibly separated 47
4.4.4 Isotonic regression degrades performance substantially . . . 48
4.5 Calibration-map complexity and calibration-set size . . . 50

4.6 Robustness . . . 52
4.6.1 Sensitivity to regularisation . . . 52
4.6.2 Sensitivity to the shot-feature specification . . . 53
4.6.3 Class-wise calibration and the draw . . . 54
4.6.4 Expected calibration error does not track probabilistic performance 57
4.7 Comparison with the market . . . 58

5 Discussion and Conclusion 59
5.1 Summary of the argument . . . 59
5.1.1 Summary of the argument . . . 59
5.2 Parametric versus non-parametric calibration . . . 59
5.3 The failure of isotonic regression . . . 61
5.4 The multiclass hypothesis . . . 61
5.5 The draw problem is a sharpness failure . . . 62
5.6 Expected calibration error is unsuitable for model selection . . . 63
5.7 The market benchmark . . . 64
5.8 Limitations . . . 64
5.9 Directions for further work . . . 65
5.10 Conclusion . . . 65

References 66

List of Figures

3.1 Optimisation history of the hyperparameter search, Random Forest (left) and XGBoost (right). 30
3.2 Estimated relative importance of each hyperparameter for validation-period logarithmic score, Random Forest (left) and XGBoost (right). 31
3.3 Parallel-coordinate representation of the Random Forest search. Each line is one trial. 31
3.4 Parallel-coordinate representation of the XGBoost search. Each line is one trial. 32

4.1 Mean bookmaker overround by year. 41
4.2 Reliability diagram for de-vigged market probabilities under multiplicative normalisation, validation and test periods combined. 42
4.3 Pairwise differences in mean per-match logarithmic score with 95% block-bootstrap intervals, XGBoost base learner. 47
4.4 Test-period logarithmic score as a function of calibration-set size, XGBoost base learner. 50
4.5 Distribution of predicted probabilities by outcome class, uncalibrated XGBoost, test period. 56
4.6 Reliability diagrams by calibration treatment, XGBoost base learner, test period. Points on the diagonal indicate agreement between predicted probabilities and observed frequencies. 56
4.7 Expected calibration error against logarithmic score for every model-calibration pipeline, test period. 57

List of Tables

3.1 Coverage of bookmaker odds and pre-match Elo ratings across all divisions in the source dataset. . . . 23
3.2 Automated tests verifying pipeline correctness. . . . 26
3.3 Feature specification, 78 variables. All are computable before kick-off; those derived from match outcomes are constructed from strictly prior fixtures. . . . 28
3.4 Hyperparameters selected by Optuna over twenty trials per model, minimising validation-period logarithmic score. . . . 30
3.5 Summary of the hyperparameter search. . . . 32

4.1 Calibration treatments ordered by fitted-map complexity (parameters and structure). . . . 38
4.2 Chronological partition of the modelling sample (six core leagues). . . . 39
4.3 1X2 outcome frequencies by league (%), modelling period. . . . 40
4.4 Out-of-sample probabilistic performance on the test period by base model and calibration method. **Bold** indicates the lowest (best) value in each column across the full table. . . . 43
4.5 Pairwise differences in mean per-match logarithmic score between calibration treatments, xgboost base learner, test period. . . . 45
4.6 Pairwise differences in mean per-match logarithmic score between calibration treatments, Random Forest base learner, test period. . . . 45
4.7 Test-period logarithmic score by calibration-set size, XGBoost base learner. . . . 51
4.8 Sensitivity of Dirichlet calibration to the strength of the L_2 penalty, XGBoost base learner, test period. . . . 53
4.9 Class-wise expected calibration error on the test period. . . . 55
4.10 Dispersion of predicted probabilities by outcome class, XGBoost base learner, test period. . . . 56

Chapter 1

Introduction

Predicting the outcomes of football matches has long been a subject of interest in sports analytics and betting research (Dixon & Coles, 1997). The large volume of historical match data, recent development of forecasting techniques using machine learning algorithms and the economic liquidity in sports betting markets make this both a scientifically challenging and economically significant problem (Bunker & Susnjak, 2022).

In the sports betting industry, bookmakers set odds using advanced statistical models and expert knowledge, and these odds, created by betting brokers and updated by the bettor's accumulated expectations, encapsulate ex-ante information about team strength, injuries, and other factors. In fact, studies have found that betting odds often outperform standalone mathematical models in predictive accuracy, after due treatment, remaining a strong baseline forecast (Spann & Skiera, 2009). This means any predictive model should be rigorously designed to approach the forecasting power inherent in market odds.

1.1 Motivation and context

Sports betting markets sit at the intersection of probabilistic forecasting and economic decision-making. For association football, the most widely available forecast are the bookmaker's odds, whose widespread availability is a direct consequence of the rapid expansion of regulated online betting, where prices are published and updated continuously across leagues (Angelini & De Angelis, 2019). These odds can be transformed into implied probabilities and interpreted as a market-based aggregate of information.

At the same time, modern open datasets provide pre-match signals of team strength and form (e.g., World Football Elo Ratings (2026) and recent results), enabling the construction of statistical and machine-learning models that estimate match outcome probabilities.

What makes a forecast useful in this setting is not classification accuracy but the quality of the estimated probabilities and the decisions derived from them (Huang, Zhao, Zhu,

Chen, & Vanden Broucke, 2020). While classification accuracy is a natural first evaluation metric, it considers only whether the most likely class prediction is correct and ignores the strength of the model's probabilistic belief. Consequently, accuracy will treat, for example, predicted probabilities of 0.51 and 0.99 identically if both exceed a given decision threshold, even though these forecasts reflect very different levels of uncertainty and should lead to different decisions (Fawcett, 2006). This reliance on hard classification decisions makes accuracy sensitive to arbitrary threshold choices and particularly misleading in the presence of class imbalance, where consistently predicting the most frequent outcome can result in deceptively high accuracy (He & Garcia, 2009).

In contrast, probabilistic forecasting explicitly evaluates the quality of predicted probabilities in terms of two distinct properties: calibration, meaning reliability - whether events predicted with a given probability occur with that frequency - and sharpness, meaning the concentration or informativeness of the predictive distribution around a single outcome, independently of whether that distribution is correct (Gneiting & Raftery, 2007). This distinction is central to what follows, as calibration and sharpness are separate properties, and a forecast can possess one without the other. A forecaster who always predicts the historical base rate is perfectly calibrated and entirely uninformative.

1.2 Problem statement

Machine-learning classifiers do not, in general, produce probability estimates that can be interpreted as such. A random forest, for instance, typically reports the fraction of trees voting for each class; because the base trees have relatively high variance (exacerbated by feature subsetting), this vote share tends to be biased away from 0 and 1, and is not, in general, well calibrated to observed frequencies (Niculescu-Mizil & Caruana, 2005a). Tree ensembles in particular are known to yield distorted probability distributions, and the standard remedy is post-hoc calibration: a transformation fitted on held-out data that maps raw model outputs onto probabilities corresponding more closely to observed frequencies.

The 1X2 problem poses a specific difficulty for this remedy. The forecast is a three-class probability vector $\mathbf{p} = (p_H, p_D, p_A)$ subject to $p_H + p_D + p_A = 1$; this constraint guarantees internal consistency of the vector, not fidelity to observed frequencies, so a normalised forecast can still be systematically miscalibrated in each component (Guo, Pleiss, Sun, & Weinberger, 2017). The two calibration methods most commonly applied in the sports forecasting literature - Platt scaling and isotonic regression - are binary techniques. Their application to a three-class problem requires fitting an independent map per outcome and renormalising the results, a corrective applied after the fact rather than a property of the methods themselves. Dirichlet calibration (Kull et al., 2019) is natively multiclass and avoids this by construction,

but it is rarely applied in football forecasting, and its performance in this setting has not been systematically assessed.

Furthermore, in a temporally ordered pipeline the calibration map must be fitted on data that is both prior to the evaluation period and representative of it, but extending the calibration window to accumulate observations relies on progressively less representative seasons, while contracting it to improve representativeness reduces the available sample.

Bookmaker odds provide the natural benchmark against which the resulting forecasts can be assessed. They embed a profit margin, such that implied probabilities sum to more than one, and may exhibit systematic pricing biases such as the favourite-longshot bias (S. Clarke, Kovalchik, & Ingram, 2017; Kuypers, 2000) and therefore require de-vigging before they can be interpreted as probabilities. A model trained to reproduce them would converge towards the market forecast and would acquire no independent information (Hubáček, Šourek, & Železný, 2019). This dissertation accordingly treats bookmaker odds as a benchmark comparator and excludes them from the model's feature set, training instead on realised outcomes and pre-match covariates.

1.3 Objectives and research questions

The objective of this research is to determine whether post-hoc calibration improves probabilistic football forecasts, which calibration method should be preferred, and why. The scope of this answer is restricted to probability quality rather than realised betting returns. The specific objectives are:

1. **Probability estimation:** Train machine-learning models, with emphasis on tree-based methods such as Random Forest and XGBoost, to estimate pre-match 1X2 probabilities from team strength and form features under strict temporal validation.

2. **Calibration comparison:** Apply and compare post-hoc calibration methods spanning a range of structures and complexities - Platt scaling and isotonic regression applied one-versus-rest, and Dirichlet calibration in both its full and diagonally constrained forms - fitted on a time-separated validation period.

3. **Probabilistic evaluation:** Assess the resulting forecasts using multiple strictly proper scoring rules and calibration diagnostics, both in aggregate and decomposed by outcome class, with statistical inference appropriate to temporally dependent data.

4. **Market comparison:** Establish de-vigged bookmaker probabilities as a benchmark, characterise their own calibration, and determine how far post-hoc calibration closes the gap between model and market.

5. **Robustness:** Determine whether the observed differences between calibration methods reflect substantive properties of those methods or artefacts of the calibration sample size and implementation.

These objectives correspond to three research questions:

- **RQ1.** Does post-hoc calibration improve the probabilistic quality of machine-learning forecasts of football 1X2 outcomes, and by how much relative to the de-vigged market benchmark?
- **RQ2.** Does the choice of calibration method materially affect the result, and if so, what property of the methods accounts for the difference - their treatment of the multiclass structure, or the complexity of the map they fit relative to the calibration sample available?
- **RQ3.** Are the observed differences robust to the size of the calibration window, to the regularisation applied, and to implementation, or are they artefacts of the experimental design?

1.4 Structure of the dissertation

The remainder of this dissertation is organised as follows:

- **Chapter 2** reviews the relevant literature on football forecasting, rating systems, machine learning for probability estimation, proper scoring rules and calibration, and the treatment of bookmaker odds as a probabilistic benchmark.
- **Chapter 3** describes the data, the temporal design, the feature construction and leakage controls, the base models, the calibration treatments and the evaluation protocol.
- **Chapter 4** reports the empirical findings.
- **Chapter 5** interprets those findings in relation to the literature, states the limitations of the design, and concludes the thesis.

Chapter 2

Literature Review

2.1 Football forecasting as a probabilistic problem and classical statistical baselines

Football outcome forecasting is best framed as a probabilistic prediction exercise. Match outcomes reflect relatively stable pre-match signals (team strength, home advantage, injuries, tactics) plus substantial irreducible uncertainty that is amplified by football's low-scoring nature. As a result, a single late goal or a red card can flip the forecasted result or the winner even when one team is better on average. For both scientific evaluation and betting applications, probability forecasts over "1X2" (a term representing a three-fold possible outcome of play, with "1" representing home victory, "X" representing a draw and "2" as the away team victory), decision quality depends on calibration and expected value, not only model accuracy, as odds - the reward a bookmaker offers for each outcome - can be interpreted as a market-based probability forecast distorted by margin and pricing effects (Hegarty & Whelan, 2025).

2.1.1 Classical baseline: goal-based Poisson models

A natural baseline is to treat goals as outcomes of a count process with match-specific scoring intensities determined by latent team strength. Empirically, conditional on an underlying "fitness" signal, goal scoring is often approximated reasonably well by Poisson processes with fixed expectations, while small departures (heterogeneity across matches, weak dependence between teams goal counts) help explain misfit around low scores and draws (Heuer, Müller, & Rubner, 2010). This is why Poisson score models remain the default classical benchmark, as they offer a tractable mapping from attack/defence strength and home advantage to probabilities for goals, scorelines, and 1X2 outcomes.

Within this family, the canonical starting point is the independent Poisson attack-defence

model with refinements aimed at two persistent issues: the dependence between teams' goal counts - most visible in 0-0, 1-0, 0-1, 1-1 and time variation in team strength (Boshnakov, Kharrat, & McHale, 2017). This model expands on the classical Dixon and Coles' model that adjusts the independent Poisson likelihood to better fit low scores and introduces time decay so recent matches weigh more when estimating strength, explicitly addressing non-stationarity in form (Boshnakov et al., 2017; Dixon & Coles, 1997). Other classical extensions (e.g., bivariate count models and dynamic latent-strength specifications) pursue the same objective: improve probability mass where it matters for the bookmakers pricing choice - especially the draw - without abandoning interpretability (Karlis & Ntzoufras, 2003; Koopman & Lit, 2015).

Even when the decision target is 1X2, modelling scores can be practically useful because it yields internally consistent outcome probabilities and connects naturally to derivative markets (exact score, totals). This "score-first, outcome-second" and the framework's predictive success helps explain why Poisson-type regression baselines remain common in the applied forecasting literature (Angelini & De Angelis, 2017; Azhari, Widyaningsih, & Lestari, 2018).

2.1.2 Alternatives that relax Poisson assumptions

Poisson processes imply memoryless inter-arrival times and variance tied to the mean, which may be adopted as much for convenience as for fit. Representative alternatives keep the count-process spirit but relax key assumptions. Boshnakov et al. (2017) propose a renewal-process approach (bivariate Weibull counts) to improve fit and evaluate forecasts using calibration and a betting criterion.

From a different angle, time-series count models incorporate dynamics directly: Angelini and De Angelis (2017) propose PARX (Poisson autoregression with exogenous predictors) as an evolution of classical baselines rather than a black-box departure, with results showing that introducing temporal dependence can materially affect the calibration of predicted score distributions over time. The main takeaway for the present study is that baseline modelling choices - such as whether match intensities are treated as static or dynamic - can influence probability estimates in outcome regions that are particularly relevant for betting decisions, such as low-scoring matches, draws, and narrow victories.

2.1.3 Baselines beyond goal models: heuristics, match stats, and crowd signals

Football forecasting also illustrates a broader methodological lesson, as more information is not automatically better. In judgmental forecasting, small signals can sometimes compete with richer but noisier knowledge - for example, the recognition heuristic where merely knowing a team can carry predictive power (Pachur & Biele, 2007). For modelling, this motivates

careful benchmarking: before claiming improvement, a method should beat strong baselines built from stable, pre-match information.

Match analysis research documents many event indicators (possession, shots, passing) related to performance Sarmento et al. (2014), and regression-based approaches can link such indicators to match success (Liu, Gómez, Lago-Peñas, & Sampaio, 2015). However, for pre-match forecasting, many event variables are unavailable at kick-off and must be used only through lagged summaries, and repeated team appearances across matches require careful treatment of dependence when modelling (Liu, Hopkins, & Gómez, 2015).

A complementary perspective is that forecasts can be generated from different information channels - ratings, crowd assessments, and market prices. For instance, Transfermarkt GmbH & Co. KG (2025) crowd valuations have been shown to produce competitive forecasts against standard rankings/ratings in some settings (Peeters, 2018). Conceptually, such work supports assessing the probability quality to check whether a new model adds information or simply re-encodes existing signals.

2.2 Elo and rating systems as team-strength signals

Team-strength rating systems summarise a club's latent ability as a time-varying score updated after each match. Unlike league rankings, which can be noisy and sensitive to unequal fixture difficulty early in the season, dynamic ratings provide a continuously updated (and relative) measure of team strength that is usable at any point in time and comparable across competitions (Hvattum & Arntzen, 2010).

The dominant approach in the field is derived from the Elo algorithm, adapted from chess to football, which maps the pre-match rating difference into an expected result (typically via a logistic transformation) and then updates each team's rating in proportion to the gap between the observed and expected outcomes (Elo, 1978). Hvattum and Arntzen (2010) describe a standard football specification that encodes outcomes as 1 (win), 0.5 (draw), and 0 (loss), with update size controlled by a learning-rate parameter (k), reflecting a trade-off between responsiveness and stability. They also discuss goal-difference scaling, where larger-margin wins induce larger updates than narrow wins.

A common use-case is to treat ratings as a compact pre-match signal that can be translated into 1X2 probabilities through a separate probabilistic layer. Hvattum and Arntzen (2010), for example, fit an ordered logit model using the home-away Elo difference as the sole predictor to obtain home/draw/away probabilities, and evaluate performance under both statistical and economic criteria - highlighting a recurring point in the literature: results-based ratings can be informative, but market odds may reflect additional information beyond what a rating captures.

Beyond baseline Elo, several "Elo relatives" adjust either the update rule or the information driving updates. Constantinou and Fenton (2013) proposes a π-rating which uses score-based discrepancies to update team strength and is evaluated against Elo variants and market odds, with reported outperformance and profitability over multiple Premier League seasons (2007/08-2011/12) after accounting for bookmaker margin. Other approaches use Elo as one component within richer forecasting structures, such as combining Elo with scoring indicators and Poisson goal models (Saribekyan & Yarovoy, 2025), often evaluated with proper scoring rules including Ranked Probability Score - which will be explored further on in this dissertation.

More recent applied work commonly uses Elo-derived variables within broader ML pipelines. Kim, Lee, and Seo (2024) include multiple Elo-related features in a large pre-match feature set and report strong dependence on model class and feature selection; Lee, Jhang, and Choi (2021) incorporate Elo difference as an opponent-strength proxy within fuzzy regression alongside venue and other contextual factors. Finally, Elo time series are also used descriptively to study performance dynamics across leagues, as in the analysis of rating trajectories and trends in Europe's top leagues by Jung and Jung (2025).

2.3 Machine learning for football probability estimation

Machine learning (ML) has become a dominant approach for forecasting football match outcomes because it can combine heterogeneous pre-match information (team-strength summaries, recent form, contextual variables, and market-derived signals) and capture non-linear interactions that are awkward to encode in fully parametric models (Hvattum & Arntzen, 2010). In betting-oriented work, however, the primary object is not the most likely class label (1X2), but a probability distribution that can be compared against odds-implied probabilities and used in expected-value decisions.

Empirical evidence suggests that reported performance is frequently driven more by data quality and feature construction than by algorithm novelty. Bunker and Susnjak (2022) emphasize that feature engineering typically dominates the incremental gains from switching among standard classifiers, and they note that many studies adopt season-ordered training and testing to reflect temporal evolution in team performance. A broader systematic review of ML in sports betting (Galekwa, Radebe, Ma, Li, & Hao, 2024) similarly documents substantial heterogeneity in data sources, leagues, targets, and evaluation practices. Together, these reviews highlight two design constraints for football probability estimation: models must be built around available pre-match signals and evaluated temporally, and cross-paper comparisons are difficult because settings differ in leagues, covariates, and objective functions.

Within applied football forecasting, tree-based ensembles (Random Forests and Gradient

Boosting variants) are recurrent because they handle mixed feature types, non-linearities, and interactions with limited manual specification. Rodrigues and Pinto (2022) compare several algorithms and evaluate both predictive metrics and betting-style returns; while they report strong overall success for their best tree model, they also show near-failure on the draw class, illustrating a common issue in 1X2 modelling: performance summaries can look good while one outcome (typically draws) is poorly learned. Chen (2025) reaches a similar conclusion in a hybrid framework combining ML with bivariate Poisson-derived inputs: multi-class prediction is systematically constrained by the difficulty of assessing draws, even when overall accuracy improves.

Because many match statistics are only known at full time, feature design is both the main differentiator and the main leakage risk. Rodrigues and Pinto (2022) explicitly avoid leakage by transforming match statistics into pre-match inputs via rolling or season-to-date aggregates computed strictly from prior games (goals, shots, cards, etc.), and they report that bookmaker odds and selected team-strength proxies are among the most informative predictors. This point is methodologically central: when odds are included as features, a model's apparent intelligence may partly reflect learning the market's forecast, which complicates later comparisons between model probabilities and market-implied probabilities.

Some ML contributions attempt richer representations beyond aggregate team summaries. Cho, Yoon, and Lee (2018) model passes as a graph (players as nodes and passes as edges), compute network metrics (e.g., centrality and connectivity), and use these features in a gradient boosting model for forecasting. They also illustrate issues that generalize to 1X2 probability modelling, like limited samples for draws, which are removed, yielding a binary win-lose task that may improve discrimination but does not solve the three-outcome probability problem required for 1X2 betting decisions.

Cross-league generalisation becomes increasingly important as datasets span many competitions and seasons. Stübinger, Mangold, and Knoll (2020) use player characteristics and match information across multiple leagues and structure evaluation around a "formation" versus "trading" period to avoid look-ahead in betting tests. Chen (2025) similarly targets robustness using multi-league data. At a broader scale, Galekwa et al. (2024) describe soccer datasets with hundreds of thousands of matches across dozens of leagues, underscoring that scale comes with distribution shift: leagues differ in goal rates, draw frequencies, competitive balance, and home advantage. Without accounting for competition structure, a single global model can be unstable, and performance can be sensitive to which leagues and seasons are included.

Overall, the literature reveals a persistent ML gap between classification and probability estimation. Some studies acknowledge this explicitly: Rodrigues and Pinto (2022) point to outcome-probability estimation as a necessary step for decision support, while Chen (2025)

uses a SoftMax output to produce normalised 1X2 probabilities.

Since betting decisions depend on well-calibrated probabilities, accuracy alone can mask systematic miscalibration - supporting the thesis emphasis on proper scoring rules and calibration as core components rather than optional add-ons for generalisation. In this line, the existing literature on calibration is reviewed in the next section.

2.4 Proper scoring rules and probabilistic evaluation and calibration

Football forecasting models must output probabilities, not just class labels, so that predictions can be compared across methods and interpreted for decision-making as signals of model confidence. This motivates evaluation frameworks that treat forecasts as predictive distributions (1X2 probability vectors) and assess them using proper scoring rules, which reward honest probabilistic forecasting and are minimized in expectation by the true predictive distribution (Gneiting & Raftery, 2007).

Probabilistic forecast quality is typically described along two dimensions: calibration, the agreement between predicted probabilities and observed frequencies - for instance, among events assigned probability 0.60, roughly 60% should occur - and sharpness, which captures how informative the forecasts are, i.e., how concentrated the predictive distribution is away from the base rate. The two are logically independent: a forecaster who always predicts the historical base rate (e.g., $p_H = 0.45$ for every home team, if that is the sample frequency of home wins) is perfectly calibrated, since predicted and observed frequencies coincide exactly, yet entirely uninformative, since every match receives the same forecast. A sharper forecaster instead varies its probabilities from match to match, issuing values closer to 0 or 1 when warranted; but sharpness is only desirable conditional on calibration, since a model can appear sharp simply by being overconfident, issuing extreme probabilities that do not match the corresponding observed frequencies, which will worsen expected decision performance and be penalized by proper scoring rules. Therefore, the goal is to achieve the greatest possible sharpness subject to good calibration (Gneiting & Raftery, 2007). Machine-learning classifiers, however, frequently fail to satisfy this calibration requirement out of the box: their raw outputs are typically ranking scores or vote shares rather than probabilities fitted to match observed frequencies, and are known to be systematically distorted in characteristic, model-dependent ways (Niculescu-Mizil & Caruana, 2005).

For football match outcomes, commonly used strictly proper scoring rules include the logarithmic score and the Brier score, the latter extending naturally to multiclass settings by comparing predicted probability vectors to realized outcomes (Brier, 1950; Gneiting & Raftery, 2007). When outcomes are viewed as ordered relative to one team (loss-draw-win),

the Ranked Probability Score (RPS) is frequently used, as it penalizes probability mass placed far from the realized category (Jordan, Krüger, & Lerch, 2019).

Because proper scoring rules combine calibration and discrimination, they are often complemented with calibration diagnostics. Reliability diagrams provide visual checks, while scalar measures such as Expected Calibration Error (ECE) summarize deviations from perfect calibration, albeit with sensitivity to binning and sample size (Posocco & Bonnefoy, 2021). These issues are particularly relevant in football due to class imbalance, especially for draws, which can affect both calibration assessment and method ranking (Huang et al., 2020).

As a result, post-hoc calibration is commonly applied as a modular step, recalibrating model outputs on held-out data. Standard approaches include Platt scaling and isotonic regression for binary tasks (Martino, De Santis, Baldini, & Rizzi, 2019), and Dirichlet calibration for multiclass problems such as 1X2 forecasting (Kull et al., 2019). Calibration robustness under distribution shift is especially important in football, where season-to-season non-stationarity and league heterogeneity are prevalent (Hekler, Brinker, & Buettner, 2023).

Overall, the literature supports treating probability evaluation and calibration as core components of football forecasting, rather than optional refinements, particularly when model outputs are compared with market-implied probabilities and used for betting-related decisions (Nikolić, Nikolić, Stefanović, Koprivica, & Stefanović, 2025).

2.5 Post-hoc probability calibration

The preceding section established calibration as a component of forecast quality distinct from sharpness, and established that machine-learning classifiers frequently fail to deliver it. This section reviews the methods available to correct that failure. Post-hoc calibration treats the problem modularly, where a model is first fitted to maximise predictive performance, and a second mapping is then estimated on held-out data to transform its outputs into probabilities that better match observed frequencies. The classifier itself is left untouched, which makes the calibration map an object that can be varied and compared independently of the model producing the scores.

The methods reviewed below differ along two dimensions that structure the comparison undertaken in this dissertation: whether the calibration map is parametric, with a fixed and small number of free parameters, or non-parametric, with flexibility that grows with the calibration sample; and whether the map is natively multiclass, operating on the full probability vector, or binary, applied one class at a time with a subsequent normalisation. The two dimensions are independent and can overlap in the same method: Platt scaling is parametric and binary, isotonic regression is non-parametric and binary, and Dirichlet calibration and vector scaling are parametric and multiclass. Examining both dimensions jointly is necessary

to explain the differences observed later in this dissertation.

2.5.1 Platt scaling

The parametric approach originates with Platt (1999), who addressed the problem that support vector machines produce real-valued decision values, f, rather than probabilities: f is the signed distance to the SVM's separating hyperplane and is unbounded, whereas a probability must lie in $[0, 1]$. Platt proposed fitting a sigmoid to these decision values, mapping f onto the probability that the class is 1 conditional on f,

$$P(y = 1 \mid f) = \frac{1}{1 + \exp(Af + B)} \tag{2.1}$$

where the two scalar parameters A and B are estimated by minimising the cross-entropy on a calibration set held out from model fitting. Rather than fitting the equation above against the hard $0/1$ labels directly, Platt replaced them with targets shrunk slightly toward 0.5, computed from the number of positive and negative calibration examples; this prevents the map from collapsing to a step function and overfitting when the model separates the calibration set perfectly. Platt also recommended replacing the hard labels with slightly shrunk targets when fitting the equation above, to prevent the map from overfitting in regions where one class dominates. The original pseudocode is prone to numerical overflow, and the stable formulation given by Lin, Lin, and Weng (2007) is the one implemented in most current software.

Platt scaling has only two free parameters, so it can be estimated from a small calibration sample. The correction it applies is necessarily sigmoid-shaped, which is appropriate when the underlying distortion has that shape and restrictive otherwise.

Niculescu-Mizil and Caruana (2005b) provide the standard empirical characterisation of when this restriction is appropriate. They document that different classifier families distort probabilities in characteristically different directions: maximum-margin methods such as boosted trees push probability mass away from zero and one, producing a sigmoid distortion that Platt scaling is well suited to invert, whereas näive Bayes pushes mass toward the extremes. Neural networks, bagged trees and random forests emerge as comparatively well calibrated before any correction is applied, and among the best after it.

Kull, Silva Filho, and Flach (2017a, 2017b) identify a further limitation. Platt scaling always forces an S-shaped correction onto the raw probabilities, even when no correction is needed: because the sigmoid has no setting of A and B that simply reproduces the input unchanged, it always makes alterations to a classifier that is already well calibrated, and ends up distorting probabilities that did not need fixing. Beta calibration uses a richer family that can reproduce the input when appropriate, and also handles the extreme-pushing distortion that logistic calibration handles poorly.

2.5.2 Isotonic regression

The non-parametric alternative was introduced to machine learning by Zadrozny and Elkan (2001, 2002). Isotonic regression assumes only that the calibration map m is monotone, and, given a calibration set (f_i, y_i) with model outputs f_i and binary labels y_i, seeks the isotonic function minimising the squared error to the labels,

$$\hat{m} = \underset{z \text{ isotonic}}{\arg\min} \sum_i (y_i - z(f_i))^2 \tag{2.2}$$

subject to $z(f_i) \leq z(f_j)$ whenever $f_i \leq f_j$. Unlike the two-parameter sigmoid in Equation 2.1, Equation 2.2 places no restriction on the functional form beyond monotonicity, so the fitted map $\hat{m}$ is a step function whose number of steps grows with the calibration sample rather than being fixed in advance. The solution is found by the Pool Adjacent Violators Algorithm of Ayer, Brunk, Ewing, Reid, and Silverman (1955): predictions are sorted by f_i, and any pair of adjacent points that violates monotonicity is pooled and replaced by the weighted average of the labels in that pool, repeated until the resulting sequence is monotone. Because it imposes no functional form beyond monotonicity, it can correct any monotone distortion rather than only sigmoid-shaped ones.

That flexibility carries a cost in data, and Niculescu-Mizil and Caruana (2005b) report that isotonic regression matches or exceeds Platt scaling only when the calibration set contains roughly a thousand or more observations, and that below this it overfits and performs worse; their companion analysis of boosted models places the crossover somewhat higher, at around two thousand. These figures are empirical heuristics from a particular experimental setting rather than general results, but the qualitative finding, that a non-parametric calibration map requires substantially more data than a two-parameter one, is consistent across the literature. Kumar, Liang, and Ma (2019) give the underlying reason in formal terms, showing that binning-based methods require a sample growing with the number of bins whereas scaling methods do not.

Moreover, because the fitted map assigns each pool the empirical frequency observed within it, it can output probabilities of exactly zero or one. Under the logarithmic score, a probability of zero assigned to an outcome that subsequently occurs incurs an infinite penalty, so isotonic outputs must be clipped away from the boundary before the logarithmic score can be computed at all. The choice of clipping threshold is therefore not a neutral implementation detail, and its effect on the reported score is examined directly in Chapter 4.

2.5.3 Extending binary calibration to three classes

Both methods above are binary. The standard extension to K classes, formalised by Zadrozny and Elkan (2002), decomposes the problem into K one-versus-rest tasks, calibrates each independently, and normalises the resulting vector to sum to one.

The deficiency of this construction is straightforward and acknowledged in the source. The K calibration maps are estimated separately and know nothing of one another, so their outputs carry no guarantee of summing to unity; the normalisation that restores coherence is applied afterwards and forms no part of any of the maps. The calibrated vector is therefore not itself the output of a coherent probability model, but a repaired collection of independent estimates. For a three-class problem in which the outcomes are mutually exclusive and exhaustive by construction, this is an awkward foundation, and it is the specific weakness that natively multiclass methods are designed to avoid.

2.5.4 Natively multiclass calibration

Two families of natively multiclass calibration map are relevant here, and both operate on the full vector.

Guo et al. (2017), addressing miscalibration in deep neural networks, define a family of maps applied to the logit vector $\mathbf{z} \in \mathbb{R}^K$, the model's raw, pre-softmax output over the K classes, before this vector is passed through the softmax function to obtain the final probability vector. Matrix scaling applies a full linear transformation,

$$\mathbf{z} \mapsto W\mathbf{z} + \mathbf{b} \tag{2.3}$$

where W is a $K \times K$ matrix and $\mathbf{b}$ is a K-dimensional intercept vector, both fitted on the calibration set, giving $K^2 + K$ free parameters in total. Vector scaling constrains W to be diagonal, so that each class k is rescaled only by its own scalar weight W_{kk} and shifted by its own intercept b_k, with no interaction between classes; this reduces the parameter count to $2K$, a scale and an intercept per class. Temperature scaling reduces this further to a single scalar T, shared across all classes,

$$\mathbf{z} \mapsto \mathbf{z}/T \tag{2.4}$$

and is a one-parameter special case of Platt scaling that cannot alter which class is deemed most probable, since dividing every logit by the same T preserves their relative ordering.

The principal finding in Guo et al. (2017) is that matrix scaling overfits on datasets with many classes, while vector and temperature scaling do not, and temperature scaling alone is frequently sufficient. The pattern is the same one Niculescu-Mizil and Caruana (2005b) report for isotonic regression against Platt scaling, arrived at in a different domain: the calibration

map is itself a model fitted to a finite sample, and its flexibility must be supported by that sample.

Kull et al. (2019) propose Dirichlet calibration, derived from the assumption that the uncalibrated probability vectors follow class-conditional Dirichlet distributions. It generalises the beta calibration of Kull et al. (2017a) to the multiclass case, and admits a convenient equivalent form: full Dirichlet calibration is a multinomial logistic regression fitted on the logarithm of the uncalibrated probability vector,

$$\mathbf{p}_{\text{cal}} = \text{softmax}(W \log \mathbf{p}_{\text{raw}} + \mathbf{b}) \,, \tag{2.5}$$

with $W \in \mathbb{R}^{K \times K}$ and $\mathbf{b} \in \mathbb{R}^K$, so that $K^2 + K$ parameters are fitted. Each calibrated probability depends on all K uncalibrated probabilities, and the softmax guarantees a valid distribution without any subsequent normalisation. Anticipating the same overfitting behaviour that Guo et al. (2017) document for matrix scaling, the authors propose ODIR, which penalises the off-diagonal elements of W and the intercepts while leaving the diagonal unregularised, shrinking the map toward a diagonal form when the calibration sample is insufficient to support the full one.

The parameter counts are worth stating explicitly for the three-class case examined here. Platt scaling and vector scaling fit six parameters, Dirichlet calibration fits twelve and isotonic regression is non-parametric. Platt scaling and vector scaling therefore have identical complexity and differ only in structure, while vector scaling and full Dirichlet calibration have identical structure and differ in complexity. This coincidence permits the two axes identified at the start of this section to be separated experimentally rather than argued about, and Chapter 3 exploits it.

2.5.5 Assessing calibration

Reliability diagrams, which plot observed frequency against predicted probability within bins, remain the standard visual diagnostic and originate in the calibration-refinement framework of DeGroot and Fienberg (1983). The corresponding scalar summary, Expected Calibration Error (ECE), was introduced by Naeini, Cooper, and Hauskrecht (2015) and computes a weighted average of the absolute gap between confidence and accuracy across bins,

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \text{acc}(B_m) - \text{conf}(B_m) \right| , \tag{2.6}$$

where the predictions are partitioned into M bins $B_1, \ldots, B_M$ by predicted probability, n is the total number of predictions, $|B_m|$ is the number of predictions falling in bin m, $\text{acc}(B_m)$ is the observed frequency of the predicted class among those predictions (the empirical accuracy

within the bin), and conf(B_m) is the average predicted probability within the bin. Each bin's contribution to ECE is therefore weighted by the share of predictions it contains, $|B_m|/n$, giving the weighted average referred to above.

ECE has since attracted a substantial critical literature. Nixon, Dusenberry, Zhang, Jerfel, and Tran (2019) show that the measure is sensitive to the binning scheme and the number of bins, and propose adaptive equal-mass binning in preference to equal-width binning. Vaicenavicius et al. (2019) demonstrate that standard estimators are biased and that apparent differences between models can be artefacts of the evaluation choices. Kumar et al. (2019) show that binned estimates systematically understate the true calibration error.

There is also a thematic cluster of criticism more fundamental and not resolved by better estimation, and it is that ECE is not a proper scoring rule - so, a forecaster can attain a low value by issuing probabilities close to the base rate, since predictions that are reliably uninformative are also reliably well calibrated. Gruber and Buettner (2022) formalise this objection and propose calibration errors derived from proper scores instead. The practical implication, and the position adopted in this dissertation, is that ECE is a useful diagnostic for locating miscalibration but an unsuitable criterion for ranking forecasters, a role that belongs to strictly proper scoring rules.

In support of this position, Murphy (1973) decomposed the Brier score into uncertainty, reliability and resolution components, and DeGroot and Fienberg (1983) separate calibration from refinement, in both cases making explicit that reliability alone is an incomplete account of forecast quality. Gneiting and Raftery (2007) state the resulting principle in the form used throughout this work - to maximise sharpness subject to calibration.

2.5.6 Calibration under distribution shift

Post-hoc calibration presumes that the relationship between predicted probabilities and observed frequencies estimated on the calibration data continues to hold on the data to which the map is applied. Where the distribution drifts, that presumption weakens. Ovadia et al. (2019) evaluate predictive uncertainty under dataset shift and find that post-hoc calibration can degrade under such shift, with temperature scaling fitted on in-distribution validation data performing worse than no calibration at all once the test distribution departs from it.

This is directly pertinent to football forecasting, where the calibration map must be estimated on one period and applied to a later one, and where team strength, tactical conventions and market pricing all evolve across seasons. It implies both that calibration should be assessed on a genuinely out-of-time test period rather than a random split, and that the size of the calibration window should be long enough to support a flexible map that extends far enough into the past to be unrepresentative of the period on which the map will be used.

2.5.7 Calibration in football forecasting

Despite the prominence of calibration in the general machine-learning literature, its application to football outcome forecasting has been limited, and applications of natively multiclass methods appear to be absent.

The closest published work is Wilkens (2026), who derives win-draw-loss probabilities from expected-goals data via a Skellam distribution and applies isotonic regression as a post-hoc calibration layer across eleven Bundesliga seasons, including a robustness check isolating the contribution of the calibration step. The method employed is binary and applied one-versus-rest.

Diniz, Izbicki, Lopes, and Salasar (2019) compare probabilistic predictive models for football using multinomial-Dirichlet constructions, but the Dirichlet there is a prior within a generative Bayesian model and calibration is assessed as a diagnostic.

Evaluation practice in the football literature is comparatively well established. Constantinou and Fenton (2012) argue for the Ranked Probability Score as the appropriate metric for 1X2 forecasts on the grounds that it respects the ordering of outcomes, although Wheatcroft (2021) disputes this, arguing that the ordering assumption is not well motivated for football and that RPS can rank forecasts implausibly. Reporting several proper scoring rules rather than relying on any one of them, as this dissertation does, is a reasonable response to that disagreement.

2.6 Betting odds as probability forecasts, de-vigging and implied probabilities

Bookmaker odds are widely used as market-based probability forecasts in football and are consistently among the most accurate publicly available predictors of match outcomes. From a forecasting perspective, odds aggregate dispersed information and therefore serve both as a benchmark for model comparison and as an object of market-efficiency analysis (Stekler, Sendor, & Verlander, 2010).

Bookmakers typically quote decimal odds, which specify the total payout (including the stake) per unit wagered if a given outcome occurs. For example, decimal odds of 2.00 correspond to receiving two monetary units for every unit staked in the event of a win. A common approximation to interpret odds as probabilities is to take their inverse, so that an outcome with decimal odds o_i is associated with an implied probability $p_i = \frac{1}{o_i}$.

However, when this inversion is applied across all mutually exclusive outcomes of a match (e.g., home win, draw, away win), the resulting implied probabilities typically sum to more than one. This excess probability mass reflects the bookmaker's built-in margin, often referred to

as the overround, which ensures profitability regardless of the match outcome. As a result, the raw implied probabilities derived from odds do not form a coherent probability distribution and cannot be directly interpreted as the market's assessment of true outcome probabilities.

To interpret odds as probabilities, the bookmaker margin must be removed through a process known as de-vigging (or margin adjustment). De-vigging rescales the implied probabilities so that they sum to one, yielding normalized probabilities that are more suitable for probabilistic comparison, calibration analysis, and expected-value calculations in betting applications (Štrumbelj, 2014). Importantly, de-vigging is not innocuous: different methods yield different probability estimates, with evidence that simple normalization can be inferior to more structured approaches and that method performance depends on market liquidity and structure (S. R. Clarke, Dixon, Pope, & Pleasance, 2017; Štrumbelj, 2014).

Probability extraction is further complicated by systematic market biases, most notably the favourite-longshot bias. In multi-outcome markets such as football 1X2, implied probabilities tend to overweight longshots relative to favourites, meaning that naive de-vigging can misrepresent how probability mass is distributed across outcomes (Newall & Cortis, 2021). Additionally, bookmaker margins are not necessarily applied uniformly across outcomes or markets, weakening the interpretation of a single overround as expected loss (Hegarty & Whelan, 2025).

From an economic perspective, market efficiency implies that extracting "fair" probabilities from odds is a measurement task, distinct from generating abnormal returns after costs. Even perfectly recovered market-implied probabilities imply negative expected value once margins are applied, and reported inefficiencies tend to be small and context-dependent (Karimov, Koshkin, Kaplun, & Butusov, 2025; Stekler et al., 2010).

Overall, the literature supports using carefully de-vigged odds as a strong baseline probability forecast, while recognizing that such probabilities are method-dependent estimates shaped by market structure and bias, and that outperforming the market is a separate and substantially harder objective (Constantinou & Fenton, 2013).

2.7 Probability quality as the decision-relevant criterion

The preceding section established bookmaker odds as a strong probabilistic benchmark. A natural further step, and the one much of the applied literature takes, is to evaluate forecasts by the returns they would have generated. This dissertation does not, and the reason is worth stating explicitly, since it determines what the remainder of the work measures.

The obstacle is not that realised return is uninteresting but that it is a poor instrument for detecting differences between forecasting methods. Returns depend jointly on the quality of the probability estimate, on the bookmaker margin, on the staking rule, and on the outcome

of whichever small subset of matches happens to fall near the decision boundary. A betting decision is a discontinuous function of a probability: a wager is placed on outcome i when $p_i o_i > 1$ and not otherwise, where p_i is the forecaster's predicted probability of outcome i and o_i is the decimal odds offered on that outcome. Two forecasters differing negligibly in probability terms can therefore differ substantially in realised return, and two differing substantially can produce identical returns over a finite sample. Realised profit is a noisy, high-variance function of the quantity actually under study.

Hubáček et al. (2019) make the related point that a model whose predictions correlate closely with bookmaker prices cannot be profitable regardless of its accuracy, since the margin absorbs any residual edge. Their proposed remedy is to seek forecasts that are informative *and* decorrelated from the market. The implication relevant here is that the market should be treated as a benchmark against which probability quality is measured, not as a target to be reproduced, and accordingly odds are excluded from the feature set in this study for exactly that reason.

Walsh and Joshi (2024) address the question directly, comparing model selection by accuracy against model selection by calibration in a betting context, and report that calibration is the more appropriate criterion because staking decisions depend on the probability estimate itself rather than on which outcome is deemed most likely. Their finding motivates the dependent variable adopted here: if probability quality is what determines decision quality, then probability quality is what should be measured, and measuring it directly avoids the additional variance that a simulated betting exercise would introduce.

This dissertation therefore evaluates forecasts using proper scoring rules and calibration diagnostics, treats the de-vigged market as the benchmark, and does not simulate returns. The economic significance of the results is framed in terms of the quality of the probabilities on which any decision rule would operate.

2.8 On temporal validation, data leakage avoidance, and non-stationarity

Although this thesis does not treat football prediction as a classical time-series forecasting problem, it remains a time-dependent supervised learning task: each observation (a match) occurs at a specific date, features must be available pre-match, and the underlying process evolves across seasons. As a result, the same core concerns emphasized in time-series evaluation - chronological validation, leakage control, and robustness under drift - still apply (Albelali & Ahmed, 2025; Falessi, Huang, Narayana, Thai, & Turhan, 2018).

Data leakage in temporally ordered pipelines is often subtle and can arise from routine steps. Forming rolling features or sequences before splitting, letting rolling windows cross

train-test boundaries, or fitting preprocessing transformations using information from future periods can substantially inflate reported performance (Albelali & Ahmed, 2025). Accordingly, the evaluation protocol must be "time-first": split chronologically before any window construction, compute rolling features strictly within each training period, and ensure that all preprocessing - scaling, imputation, calibration, and hyperparameter selection - is fitted only on past data and then carried forward (Albelali & Ahmed, 2025).

Even if leakage is avoided, the validation design can still distort conclusions. Random k-fold cross-validation assumes exchangeability and can mix seasons in a way that is inconsistent with deployment, overstating generalization when the data-generating process changes over time (Falessi et al., 2018). Moreover, different temporal evaluation schemes (e.g., walk-forward and sliding windows) can yield meaningfully different performance estimates and even change which model appears best, so robustness should be assessed under realistic time-respecting protocols rather than a single split (Hespeler, Moriano, Li, & Hollifield, 2025).

Finally, football forecasting is affected by non-stationarity: team strength, tactics, league environments, and market behaviour drift over time, inducing temporal distribution shifts and concept drift that can degrade deployment performance relative to retrospective estimates (Zhao, Liu, & Prakash, 2025). In betting-oriented modelling, this interacts with strategy iteration risk: trying many feature sets, thresholds, and model variants can produce apparently strong backtests by chance. This motivates strict separation of training, tuning, and final testing, and conservative out-of-sample testing designs to reduce backtest overfitting (Arian, Norouzi Mobarekeh, & Seco, 2024).

Chapter 3

Methodology

This chapter describes the empirical pipeline, which proceeds in six stages: data sourcing and cleaning, chronological partitioning, feature construction, base-model training, post-hoc calibration, and evaluation. The organising constraint throughout is temporal: every metric used to describe a match is computable before that match is played, and every fitted object is estimated on data strictly prior to the period on which it is assessed.

3.1 Data

3.1.1 Source

The analysis uses the dataset *Club Football Match Data (2000-2025)* compiled by adamgbor (2025), which aggregates match-level records from 38 divisions between July 2000 and June 2025. The dataset comprises 230,557 matches and 48 fields, including match date, competition, participating clubs, half-time and full-time scores, in-match statistics where available, pre-match club Elo ratings, and bookmaker odds for the 1X2, over/under and handicap markets.

Every record on the dataset is timestamped, permitting strict chronological partitioning, and the pre-match Elo ratings are supplied as of the date of each fixture, so they constitute genuinely ex-ante information rather than a retrospective reconstruction.

3.1.2 Competition scope

Coverage is not uniform across the 38 divisions. Elo ratings and bookmaker odds are available for between 98% and 100% of matches in the principal European competitions, whereas several divisions - including the lower English tiers and a number of non-European leagues - have effectively no Elo coverage, reflecting gaps in the mapping between club identifiers in the source tables. Including these would introduce systematic missingness in the

covariate that the literature identifies as the most informative pre-match signal (Hvattum & Arntzen, 2010).

The analysis is therefore restricted to six core competitions: the English Premier League, Spanish La Liga, Italian Serie A, German Bundesliga, French Ligue 1 and Portuguese Primeira Liga. This yields 50,334 matches. The restriction is applied after feature development, for reasons set out in Section 3.4.

Table 3.1 reports the availability of bookmaker odds and pre-match Elo ratings across all 38 divisions in the source dataset, and identifies the six competitions retained for analysis. The divisions exhibiting negligible Elo coverage do so because of gaps in the mapping between club identifiers in the match and rating tables, rather than because rating histories are unavailable for those clubs.

Table 3.1: Coverage of bookmaker odds and pre-match Elo ratings across all divisions in the source dataset.

Division	Matches	Odds (%)	Elo (%)	Retained
E1	13,606	98.8	99.0	
E2	13,220	98.4	12.3	
E3	12,777	98.7	0.0	
SP2	10,497	96.4	96.8	
I2	9,921	96.8	99.8	
E0	9,410	99.1	100.0	✓
I1	9,012	98.7	100.0	✓
SP1	9,008	99.0	100.0	✓
F1	8,756	99.0	99.8	✓
F2	8,677	95.7	98.9	
D1	7,522	99.0	100.0	✓
T1	7,458	96.6	92.9	
N1	7,288	98.8	96.1	
D2	7,161	98.8	99.6	
EC	7,134	99.5	0.0	
P1	6,626	98.6	99.5	✓
B1	6,559	98.6	94.3	
ARG	5,330	100.0	0.0	
USA	5,233	100.0	0.0	
SC0	5,124	98.9	95.7	
BRA	4,850	100.0	0.0	
G1	4,128	97.9	95.6	
SC1	4,094	98.7	11.7	
SC2	4,092	98.7	0.0	
JAP	4,089	100.0	0.0	
MEX	4,080	100.0	0.0	
SC3	4,010	98.6	0.4	
ROM	3,640	100.0	22.4	
POL	3,567	100.0	39.5	
SWE	3,110	100.0	52.9	
NOR	3,092	100.0	82.5	
RUS	3,008	99.9	27.6	
DEN	2,634	100.0	76.3	
CHN	2,544	99.8	0.0	
IRL	2,367	100.0	0.0	
FIN	2,354	100.0	18.4	
AUT	2,306	100.0	55.7	
SUI	2,273	100.0	0.0	

3.1.3 Cleaning

Records with a missing or invalid full-time result are removed; three such rows exist across the full dataset. Matches lacking complete 1X2 odds are retained in the frame but flagged,

since they remain valid for the purpose of accumulating club history while being unusable for market comparison. Records are sorted chronologically before any rolling computation.

3.2 Chronological partitioning

The sample is divided into four consecutive, non-overlapping periods.

- **Burn-in**, to 30 June 2005 (9,124 matches). Used solely to seed the rolling form and streak histories - the trailing-window features (such as points won or goals scored over a team's last n matches) and win/loss/draw streak counters that feed later models, detailed in Section 4 - so that these features are populated from the first match of the Training period onward, rather than starting from an empty history. No burn-in match enters model fitting or evaluation.
- **Training**, 1 July 2005 to 30 June 2021 (32,831 matches). Used to fit the base models.
- **Validation**, 1 July 2021 to 30 June 2023 (4,264 matches). Used to select hyperparameters and, subsequently, to fit the calibration maps.
- **Test**, 1 July 2023 to the end of the sample (4,115 matches). Held out entirely and used once, for final evaluation.

The length of the burn-in period exceeds what the rolling windows alone require, which need at most ten prior matches, or a few months, to be fully populated. It is set instead by the reliability of the market benchmark in the early part of the sample, as it has the steepest decline occurring between 2003 and 2008. A burn-in period ending mid-2005 does not eliminate this trend, but it excludes the years in which the bookmaker margin was highest and least representative of current pricing, at the cost of training matches that are in any case the least comparable to the evaluation period. The cut-off is a balance between this consideration and the value of additional training data, rather than a value derived from a single criterion, as extending the burn-in period further into the decline would trade an already-diminishing benefit against a fixed and certain cost in training observations.

The validation period serves two purposes in sequence: hyperparameter selection first, then calibration fitting. By the time the calibrators are fitted, it has therefore already been used once, for hyperparameter selection. What matters for the validity of the design is that it remains strictly prior to the test period, which is never used to fit any model, hyperparameter, or calibration map.

3.3 Leakage control

Data leakage in temporally ordered pipelines produces improved metrics rather than an error, so the following controls are applied:

- **Chronological ordering precedes all window construction**, and rolling features are computed with the current match excluded. Concretely, for any metric engineered from a match outcome, the sequence is: sort by date, group by club, shift by one match, then apply the rolling window. The shift is what excludes a match's own result from its own features.
- **Pre-match Elo ratings are exempt from this shift**, since they already represent the rating a club carries into the fixture. Elo momentum, being the difference between two such pre-match ratings, is likewise exempt.
- **Preprocessing transformations are fitted on the training period alone** and applied unchanged thereafter, with imputation the relevant case here. Hyperparameter selection uses the validation period only; no cross-validation scheme mixes observations across the chronological boundaries.

Then, two independent checks verify that these controls hold in practice:

- **Cross-check against an independent implementation.** The source data ships its own rolling form variables, computed by the dataset's author. Recomputing those variables from scratch and comparing reproduces the supplied values on 41,193 of the 41,202 modelled matches. The nine disagreements all occur in a club's opening fixtures, where the window is partial and the two implementations differ in how they handle an incomplete history.
- **Automated test package.** These properties are verified by tests rather than by inspection. The decisive test alters the score of a future match, recomputes the entire feature set, and asserts that every feature of every earlier match is unchanged. A feature that reads forward in time would fail this test; the implementation passes it.

Table 3.2: Automated tests verifying pipeline correctness.

Module	What it verifies	Tests
data_loader	Raw data loads with the expected schema; chronological splits do not overlap; season/date parsing is correct	7
features	Rolling features use only prior matches; the leakage guard (altering a future match does not change past features); edge cases at the start of a club's history	10
models	Baseline and model outputs have the expected shape and range; predictions are reproducible under a fixed seed	-
calibration	Calibration maps preserve valid probability distributions; Platt/isotonic/Dirichlet/vector-scaling behave correctly on synthetic inputs with known calibration error	-

3.4 Feature Engineering

3.4.1 Construction order

Features are computed on the full 38-division sample and the six-league restriction is applied only afterwards. The reason is that clubs move between divisions: a side relegated from a core league continues to play, and computing rolling windows only within the core leagues would return that club on promotion with an empty history. If a club is relegated from one of these leagues, its matches in the lower division are absent from the dataset entirely, since that competition is not among the six, and this holds even if the club is later promoted back. Feature history, however, follows the club rather than the competition: on return to a sampled league, a club's rolling form and streak counters resume from its last appearance in the sample, rather than restarting from an empty history, so the rolling window can span a gap containing no observed matches.

To make this tractable, the match table is reshaped into long format - one row per club per match - before any rolling computation, and pivoted back to wide format afterwards. Each feature is therefore a single grouped operation over club history rather than parallel home and away code paths that could diverge.

3.4.2 Feature set

The final specification comprises 78 features in four families.

Team strength. Home and away pre-match Elo ratings, their difference, and Elo momentum computed as the change in a club's own rating over its previous five and ten matches.

Form. Points accumulated over the previous three, five and ten matches; rolling means of goals scored, goals conceded and goal difference over the same windows.

Schedule. Days of rest since the club's previous fixture, and the number of matches played in the preceding fourteen days as a measure of congestion.

Streaks. Length of the current unbeaten, winless, winning, losing and drawing run entering the fixture.

Home-minus-away differentials are supplied for the principal quantities. Although tree ensembles can in principle recover a difference from two levels, providing it directly is inexpensive, and the Elo difference in particular is the single covariate on which the classical specification of Hvattum and Arntzen (2010) relies entirely.

Table 3.3 lists the complete set of covariates supplied to the models. All are computable before kick-off, and every variable derived from a match outcome is constructed from strictly prior fixtures, as described in Section 3.3.

Variable names follow a fixed convention - the suffix _home, _away, or _diff indicates the home team's value, the away team's value, or the home-minus-away difference, respectively, and a numeral in the name (3, 5, or 10) gives the rolling window, in matches, over which the quantity is averaged; for example, mean10 is the mean over a team's preceding 10 matches. The variable stems abbreviate as follows: gf, ga, and gd denote goals for, goals against, and goal difference; sot_for and sot_against denote shots on target taken and conceded; shots_for and shots_against denote total shots taken and conceded; form denotes points won; elo_momentum denotes the change in Elo rating over the given window; congestion_14d denotes the number of matches played in the preceding 14 days; and rest_days denotes days elapsed since the team's previous league match.

Table 3.3: Feature specification, 78 variables. All are computable before kick-off; those derived from match outcomes are constructed from strictly prior fixtures.

Family	n	Variables
Team strength (Elo)	3	elo_away_raw, elo_diff, elo_home_raw
Elo momentum	6	elo_momentum10_away, elo_momentum10_diff, elo_momentum10_home, elo_momentum5_away, elo_momentum5_diff, elo_momentum5_home
Form (points)	9	form10_away, form10_diff, form10_home, form3_away, form3_diff, form3_home, form5_away, form5_diff, form5_home
Rolling goals	21	ga_mean10_away, ga_mean10_home, ga_mean3_away, ga_mean3_home, ga_mean5_away, ga_mean5_home, gd_mean10_away, gd_mean10_diff, gd_mean10_home, gd_mean3_away, gd_mean3_diff, gd_mean3_home, gd_mean5_away, gd_mean5_diff, gd_mean5_home, gf_mean10_away, gf_mean10_home, gf_mean3_away, gf_mean3_home, gf_mean5_away, gf_mean5_home
Schedule	6	congestion_14d_away, congestion_14d_diff, congestion_14d_home, rest_days_away, rest_days_diff, rest_days_home
Streaks	6	draw_streak_away, draw_streak_home, loss_streak_away, loss_streak_home, win_streak_away, win_streak_home
Other	27	shots_against_mean10_away, shots_against_mean10_home, shots_against_mean3_away, shots_against_mean3_home, shots_against_mean5_away, shots_against_mean5_home, shots_for_mean10_away, shots_for_mean10_home, shots_for_mean3_away, shots_for_mean3_home, shots_for_mean5_away, shots_for_mean5_home, sot_against_mean10_away, sot_against_mean10_home, sot_against_mean3_away, sot_against_mean3_home, sot_against_mean5_away, sot_against_mean5_home, sot_for_mean10_away, sot_for_mean10_diff, sot_for_mean10_home, sot_for_mean3_away, sot_for_mean3_diff, sot_for_mean3_home, sot_for_mean5_away, sot_for_mean5_diff, sot_for_mean5_home

3.4.3 Excluded features

One class of variables is deliberately excluded, and a second class is retained but raises data-coverage issues that require comment.

Bookmaker odds are not model inputs. Training on odds would cause the model to reproduce the market forecast, eliminating the informational independence required for a meaningful comparison between the two (Hubáček et al., 2019). Odds enter the analysis only as the de-vigged benchmark.

Shot statistics are retained, but their coverage is uneven and the consequences require

comment. Fewer than 0.5% of modelling rows lack shot data in five of the six competitions; in the Portuguese Primeira Liga the figure is 54.2%, because the source records no shot statistics for that competition before the 2017/18 season. The gap is structural rather than random, and it is confined to the earlier part of the sample.

The missingness is asymmetrical, as it it affects the training period only - the validation window on which the calibration maps are fitted (2021-2023) and the test window on which they are evaluated (2023-2025) both have complete coverage, so no calibration map is estimated from imputed values and no reported score is computed from them. Moreover, missingness is handled by mechanisms appropriate to each learner, with XGBoost assigning a default split direction and the Random Forest pipeline imputing from training-period medians. Furthermore, every calibration treatment operates on an identical set of base-model outputs. Any deficiency induced in the base learner by incomplete covariates is therefore common to all calibration specifications in this dissertation and cannot bias the comparison between them, which is the central comparison of this work.

3.5 Base models

The prediction target is the 1X2 outcome (1 = home win, X = draw, 2 = away win), treated as a three-class problem with classes ordered home, draw, away. This role-based convention is fixed across matches, unlike a team-relative "win/draw/loss" label, which would require specifying a team. All models output a full probability vector rather than a class label.

3.5.1 Baselines

Two baselines are estimated: the base-rate forecast assigns the training-period class frequencies to every match, establishing the floor below which no forecast should fall. The Elo logistic regression is a multinomial logit on the pre-match Elo difference alone, following Hvattum and Arntzen (2010), and establishes what is obtainable from team strength without further feature engineering.

3.5.2 Machine-learning models and hyperparameter tuning

Two tree ensembles are used: a Random Forest, with probabilities obtained by averaging leaf frequencies, and XGBoost under the multi:softprob objective, which optimises the multiclass logarithmic score directly.

Hyperparameters are selected with Optuna over twenty trials per model, by Bayesian optimisation using the tree-structured Parzen estimator. The tuning objective is validation-period logarithmic score, not accuracy, since the thesis concerns probability quality, and selecting

models on a criterion that ignores probability magnitude would be inconsistent with its aims. The selected configurations, reported in Table 3.4, are refitted on the training period alone; the validation period is not used for fitting at any stage, being reserved for the calibration maps.

Table 3.4: Hyperparameters selected by Optuna over twenty trials per model, minimising validation-period logarithmic score.

Model	Parameter	Value
Random Forest	n_estimators	500
	max_depth	8
	min_samples_leaf	19
	max_features	0.5
XGBoost	n_estimators	100
	max_depth	5
	learning_rate	0.0425
	subsample	0.6842
	colsample_bytree	0.8343
	min_child_weight	9
	reg_lambda	1.2194

Figure 3.1 shows the objective value attained at each trial together with the running best. Both searches converge well before the twentieth trial, indicating that the trial budget was adequate and that the reported configurations are not an artefact of premature stopping.

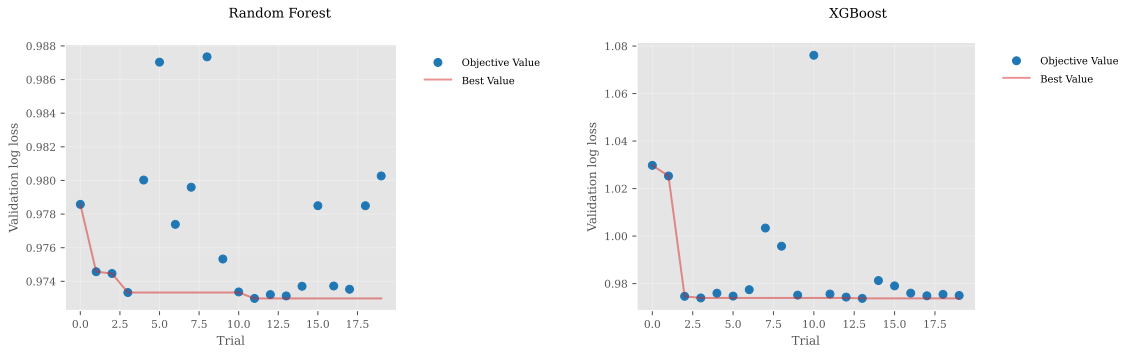


Figure 3.1: Optimisation history of the hyperparameter search, Random Forest (left) and XGBoost (right).

Figure 3.2 reports the relative influence of each hyperparameter on the objective, estimated across completed trials. This is informative beyond the selected values themselves, since it identifies which choices materially affected performance and which were largely immaterial.

Figures 3.3 and 3.4 represent each trial as a line traversing the hyperparameter axes, showing the regions of the search space on which the sampler concentrated as it converged.

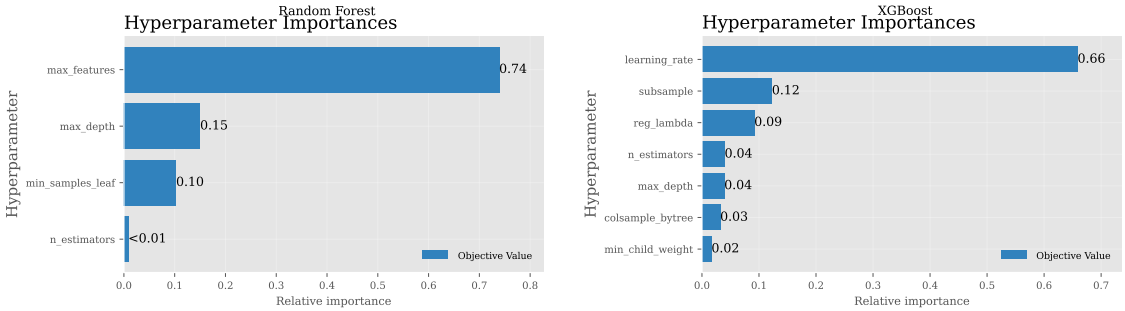


Figure 3.2: Estimated relative importance of each hyperparameter for validation-period logarithmic score, Random Forest (left) and XGBoost (right).

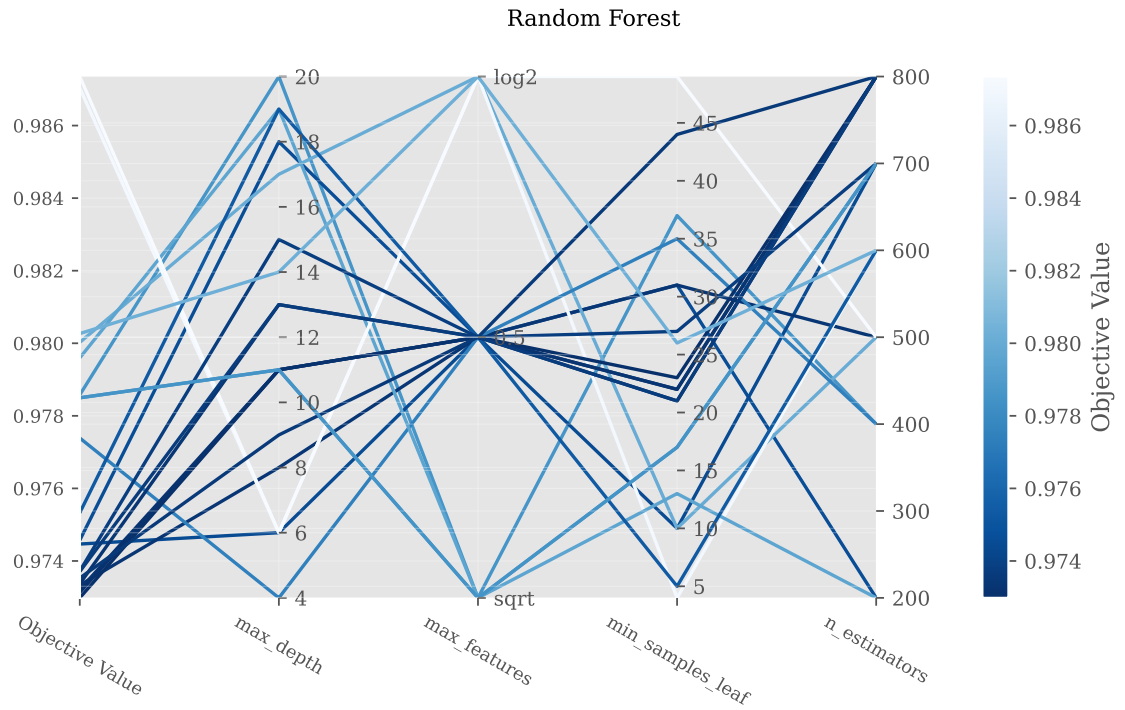


Figure 3.3: Parallel-coordinate representation of the Random Forest search. Each line is one trial.

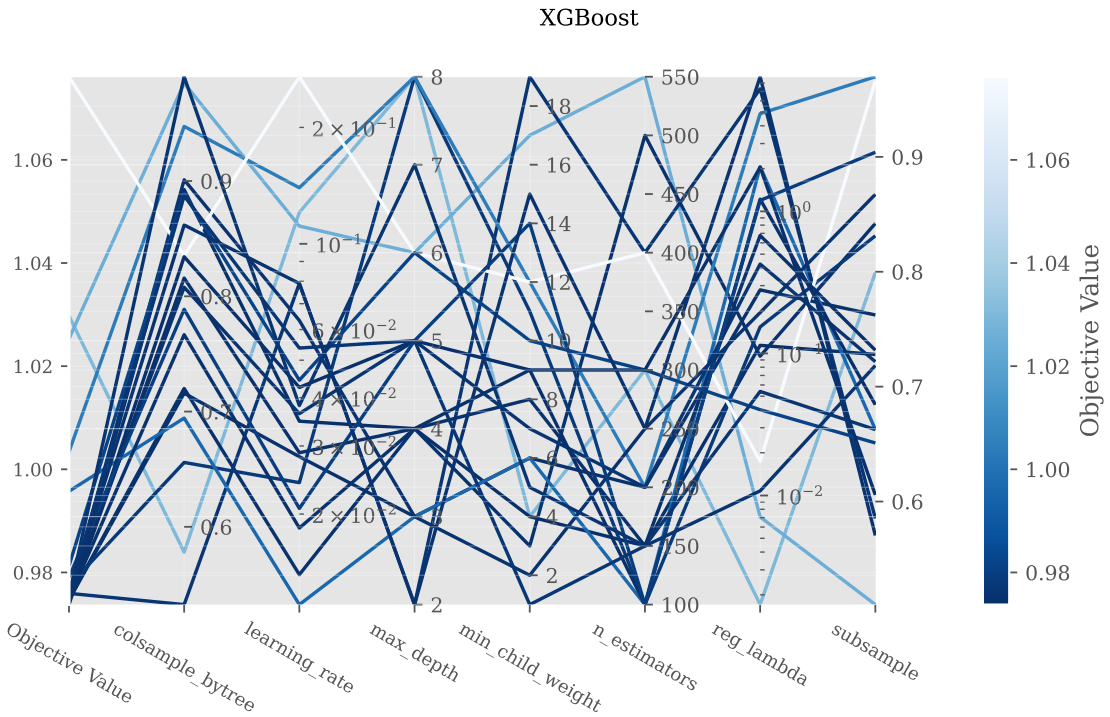


Figure 3.4: Parallel-coordinate representation of the XGBoost search. Each line is one trial.

Table 3.5 summarises the dispersion of objective values across the search, and one comparison in it deserves a direct answer. The spread between the best and worst configurations is 0.0144 for the Random Forest and 0.1024 for XGBoost, in both cases larger than any difference between calibration methods that will be later reported in Chapter 4. The two quantities are not, however, commensurable. The tuning spread is measured across different models on validation data and includes configurations that were never candidates for selection - XGBoost's worst trial uses a learning rate of 0.27 - whereas the calibration comparisons of Chapter 4 are paired on identical base-model outputs and assessed by a bootstrap that respects the temporal dependence of the test period. These two exercises measure different things: the tuning spread reflects how much performance varies across deliberately mismatched configurations, while the calibration comparisons measure a difference between two treatments applied to the same, fixed predictions. A large tuning spread says nothing about how precisely a calibration difference can be estimated.

Table 3.5: Summary of the hyperparameter search.

Model	Trials	Best	Median	Worst	Spread
Random Forest	20	0.9730	0.9749	0.9874	0.0144
XGBoost	20	0.9738	0.9758	1.0761	0.1024

3.6 Post-hoc calibration

Calibration maps are fitted on the validation period and applied to the test period. Five treatments are compared, distinguished by the structure and complexity of the map.

No calibration

The raw model output serves as the control. It is the identity map applied to the base learner's probabilities, against which every calibration treatment below is measured.

Platt scaling

Platt scaling is applied one-versus-rest: a logistic map is fitted independently to each outcome's log-odds, and the three resulting outputs are renormalised to sum to unity. This fits six parameters in total. Platt scaling is fundamentally a binary method, and the renormalisation step is a corrective applied after the fact rather than a property of the method itself.

Isotonic regression

Isotonic regression is likewise applied one-versus-rest: a monotone step function is fitted per outcome, and the outputs are renormalised in the same way as for Platt scaling. Because it is non-parametric, its effective complexity is unbounded and grows with the calibration sample rather than being fixed in advance.

Because isotonic regression can assign probabilities of exactly zero, all calibrated probabilities are floored at 10^{-12} and renormalised before evaluation. This threshold is a modelling choice rather than one dictated by any external convention: it is chosen to be small enough not to materially distort the probability mass, while avoiding the $-\infty$ penalty that an exact zero would incur under the logarithmic score. It is applied uniformly across all five specifications, and the sensitivity of the isotonic results to this choice - including the effect of an alternative, tighter floor of 10^{-15} - is reported explicitly in the robustness analysis.

Dirichlet calibration

Dirichlet calibration is natively multiclass rather than adapted from a binary method. Following Kull et al. (2019), full Dirichlet calibration is equivalent to a multinomial logistic regression on the logarithm of the raw probability vector,

$$\mathbf{p}_{\text{cal}} = \text{softmax}(W \log \mathbf{p}_{\text{raw}} + \mathbf{b}) \,, \tag{3.1}$$

with $W \in \mathbb{R}^{3\times 3}$ and $\mathbf{b} \in \mathbb{R}^3$, twelve parameters in total. Each calibrated probability depends on all three raw probabilities, and the softmax guarantees a valid distribution without

any renormalisation step. Regularisation is applied as a standard L_2 penalty on all coefficients. This approximates but does not reproduce the off-diagonal and intercept regularisation (ODIR) recommended by Kull et al. (2019); the sensitivity of results to this choice is examined directly in the robustness analysis.

Vector scaling

Vector scaling is Equation 3.1 with W constrained to be diagonal: three scale parameters and three intercepts, six in total. Its inclusion is a deliberate design choice rather than an additional method for its own sake. Comparing Platt scaling directly against full Dirichlet calibration confounds two properties, since the two treatments differ simultaneously in structure (binary-plus-renormalisation versus native softmax) and in complexity: Platt scaling has two free parameters, A and B in Equation 2.1, fitted separately per class and renormalised, whereas full Dirichlet calibration has $K^2 + K = 12$ free parameters for $K = 3$ classes, the full 3×3 matrix W plus a three-dimensional intercept $\mathbf{b}$.

Vector scaling resolves this, as it is natively multiclass, like Dirichlet, but its six parameters place it structurally between the two, equal in count to Platt's binary-plus-renormalisation total across the three one-versus-rest fits than to full Dirichlet's twelve. The contrast between Platt scaling and vector scaling therefore isolates the effect of structure at comparable complexity, and the contrast between vector scaling and full Dirichlet calibration isolates the effect of complexity (six against twelve parameters) at fixed structure.

3.7 Market baseline

Bookmaker decimal odds o_i are converted to implied probabilities by inversion, $q_i = 1/o_i$ for each outcome $i \in \{H, D, A\}$. Because the odds are set to guarantee a margin for the bookmaker, these raw inverses sum to more than unity,

$$\sum_i q_i = 1 + \nu \tag{3.2}$$

where $\nu > 0$ is the overround (the bookmaker's margin, typically a few percent). The simplest correction, multiplicative normalisation, rescales each q_i by the same factor so that the result sums to one:

$$p_i = \frac{q_i}{\sum_j q_j} \tag{3.3}$$

This spreads the margin across all three outcomes in proportion to their raw inverse odds, which implicitly assumes the bookmaker's margin is loaded equally, in relative terms, onto favourites and longshots alike. Empirically this is not the case, as bookmakers tend to shorten favourites' odds by more, and lengthen longshots' odds by more than proportionality would

predict, a pattern known as the favourite-longshot bias. The power method of Štrumbelj (2014) corrects for this by raising each q_i to a common power before normalising,

$$p_i = \frac{q_i^{1/k}}{\sum_j q_j^{1/k}} \tag{3.4}$$

where the exponent $k > 1$ is chosen, separately for each match, as the value that makes the resulting probabilities sum exactly to one; since Equation 3.4 has no closed-form solution for k, it is found numerically by root-finding. Raising each q_i to the power $1/k$ shrinks larger values (favourites) proportionally less than smaller values (longshots) before normalisation, correcting the bias in a way that the uniform rescaling of Equation 3.3 cannot.

Single-bookmaker prices are used throughout. Best-of-market prices are also present in the data and carry a substantially smaller margin, but they are recorded only from 2005 and intermittently thereafter, and they presume a bettor holding accounts with numerous operators that all remain unrestricted. The conservative choice is adopted.

3.8 Evaluation

3.8.1 Proper scoring rules

Three strictly proper scoring rules are reported, each evaluated on a match's predicted probability vector $\mathbf{p} = (p_H, p_D, p_A)$ against its realised outcome, encoded as a one-hot vector $\mathbf{a} = (a_H, a_D, a_A)$ with $a_i = 1$ for the outcome that occurred and 0 otherwise.

- **Logarithmic score.** The negative log probability assigned to the realised outcome,

$$\text{LogS} = -\log p_{i^*} \tag{3.5}$$

  where i^* is the outcome that occurred (so $p_{i^*} = p_H$, p_D, or p_A as applicable). It is unbounded and diverges as an assigned probability approaches zero for an outcome that occurs, since $-\log(0) = \infty$.

- **Multiclass Brier score.** The squared error of the probability vector against the one-hot outcome, summed over classes:

$$\text{Brier} = \sum_{i \in \{H,D,A\}} (p_i - a_i)^2 \tag{3.6}$$

  It is bounded, between 0 (a certain, correct forecast) and 2 (a certain, incorrect forecast), and penalises error quadratically rather than logarithmically, so it does not diverge when a probability of zero is assigned to an outcome that occurs.

- **Ranked Probability Score.** Treats outcomes as ordered, comparing cumulative distributions rather than the outcome-wise probabilities directly. With outcomes ordered $1, \ldots, r$ (here $r = 3$), it is defined as

$$\text{RPS} = \frac{1}{r-1} \sum_{k=1}^{r-1} \left(\sum_{i=1}^{k} p_i - \sum_{i=1}^{k} a_i \right)^2 \tag{3.7}$$

the average squared difference between the predicted and observed cumulative distributions, taken over all cut points k short of the final, trivial one where both cumulative sums equal 1. Because it compares cumulative sums rather than individual probabilities, RPS penalises probability mass placed far from the realised category more heavily than mass placed on an adjacent category - for instance, predicting a draw when the true outcome is a home win is penalised less than predicting an away win would be, since the draw's cumulative sum is closer to the home-win outcome's. The ordering used throughout is away win $<$ draw $<$ home win, which places the draw between the two win outcomes; RPS is not invariant to this choice, and reversing it would change the reported values.

3.8.2 Calibration diagnostics

Reliability diagrams and expected calibration error are reported, both in aggregate and decomposed by outcome class. ECE is treated as a diagnostic rather than as a criterion for ranking forecasters as it is not a proper scoring rule: a forecaster issuing probabilities near the base rate can attain a low ECE while conveying no information. Class-wise decomposition is reported because aggregate measures can conceal miscalibration of the minority draw class.

3.8.3 Statistical inference

Differences in mean score between pipelines are assessed by a block bootstrap (Künsch, 1989). A paired test on per-match score differences would assume independence, which does not hold: clubs recur throughout the test period, form persists, and the test window is a single realisation of a non-stationary process and a test would understate standard errors.

The procedure resamples contiguous calendar months with replacement, preserving dependence within blocks while randomising across them, and reports percentile confidence intervals for the mean paired difference in logarithmic score. The test period spans 21 months, and 10,000 resamples are drawn.

Two contrasts were specified in advance as the tests of the study's hypotheses: Platt scaling against vector scaling, which varies calibration-map structure at fixed parameter count, and vector scaling against full Dirichlet calibration, which varies parameter count at fixed

structure. These are the comparisons the design of Section 3.6 was constructed to make interpretable. The remaining pairwise comparisons are reported for completeness and are descriptive.

3.9 Robustness analysis

Because several anticipated findings are negative, and negative findings are not distinguishable from implementation error without explicit checks, three robustness analyses are specified in advance.

A calibration learning curve refits every calibrator on random subsamples of the validation period, from 250 matches to the full 4,264, and evaluates each on the unchanged test period, with twenty draws per size. If the relative performance of methods depends on calibration-set size, this will reveal it.

A regularisation sweep varies the L_2 penalty applied to Dirichlet calibration across four orders of magnitude, testing whether any null result for that method is attributable to the choice of penalty strength.

A specification check repeats the pipeline with the shot-derived features removed, retraining the base models and refitting every calibration map, testing whether the pattern of results depends on the covariate family whose coverage is uneven (Section 3.4.3).

3.10 Software and reproducibility

The implementation is in Python. Data handling uses pandas and numpy; models and calibration use scikit-learn and xgboost; hyperparameter search uses optuna; numerical optimisation for vector scaling uses scipy; figures use matplotlib.

Random seeds are fixed throughout the pipeline, and the leakage and correctness properties described in Section 3.3 are enforced by tests covering the data loader, the feature construction, the calibration implementations and the evaluation metrics.

Chapter 4

Results

4.1 Overview

Consistent with the framing established in Chapter 3, the object of study is not the predicted class label but the complete three-class probability vector $\mathbf{p} = (p_H, p_D, p_A)$, and the primary experimental factor is the post-hoc calibration method applied to that vector. Random Forest and XGBoost function as base learners generating raw probabilities; the analysis concentrates on what occurs after the base learner, once a calibration map has been applied.

The organising variable of the analysis is the complexity of that map, measured by the number of free parameters it fits. Five treatments are compared, ordered by complexity in Table 4.1. Section 4.5 shows that complexity accounts for essentially the whole pattern of results, and that the two findings which might otherwise appear unrelated - the failure of isotonic regression and the absence of any substantial advantage for full Dirichlet calibration - are manifestations of a single underlying constraint.

Table 4.1: Calibration treatments ordered by fitted-map complexity (parameters and structure).

Treatment	Structure	Parameters
Raw	None (identity map)	0
Platt (OvR)	Binary, one map per class, renormalised	6
Vector scaling	Multiclass softmax, diagonal W	6
Dirichlet (full)	Multiclass softmax, full W	12
Isotonic (OvR)	Binary, non-parametric, renormalised	Unbounded

Section 4.2 describes the modelling sample and characterises the market baseline. Section 4.3 reports uncalibrated base-learner performance. Section 4.4 presents the calibration comparison, and Section 4.5 develops the complexity account. Section 4.6 reports the checks that discriminate between substantive findings and artefacts of particular analytical choices.

Section 4.7 situates all forecasts relative to the market.

4.2 The modelling sample and the market baseline

4.2.1 Sample construction

The source dataset comprises 230,557 matches across 38 divisions between July 2000 and June 2025. As introduced in Section 3, the analysis is restricted to six core European competitions - the English Premier League, Spanish La Liga, Italian Serie A, German Bundesliga, French Ligue 1 and Portuguese Primeira Liga - yielding 50,334 matches; this restriction is repeated here in summary form as a reminder before the results that follow. It is not arbitrary. Coverage of the pre-match Elo ratings and of the bookmaker odds lies between 98% and 100% for each of these six competitions, whereas several other divisions in the source data, including the lower English tiers and a number of non-European leagues, exhibit effectively no Elo coverage at all. Their inclusion would introduce systematic missingness in precisely the feature that the literature identifies as the single most informative pre-match signal (Hvattum & Arntzen, 2010).

Table 4.2 reports the resulting chronological partition. The burn-in period exists so that no modelled match is described by rolling windows computed from an incomplete history; it seeds the form and Elo features and is excluded from all model fitting and evaluation. The effective modelling sample comprises 41,210 matches, of which 32,831 are used for training, 4,264 for validation and hyperparameter selection, and 4,115 are held out as the test period.

Table 4.2: Chronological partition of the modelling sample (six core leagues).

Split	Matches	From	To	Seasons
Burn-in	9,124	2000-07-28	2005-05-29	5
Train	32,831	2005-07-29	2021-05-23	16
Validation	4,264	2021-08-06	2023-06-04	2
Test	4,115	2023-08-11	2025-05-25	2
Total	50,334			

4.2.2 Outcome frequencies

Table 4.3 reports the 1X2 outcome frequencies over the modelling period. Home wins occur in 45.1% of matches, away wins in 29.3% and draws in 25.6%. The home advantage is substantial and persistent, which entails that a forecaster predicting a home win in every match would attain an accuracy of approximately 46% while conveying no probabilistic information at all - an illustration of the inadequacy of accuracy as an evaluation criterion, as argued in

Chapter 1. The draw, meanwhile, is the least frequent outcome and its frequency is notably stable across the six leagues. The draw is consequently both the minority class and, as the literature consistently documents, the outcome that classifiers handle least well. It is examined separately in Section 4.6.3, where the conventional interpretation of that difficulty is analysed.

Table 4.3: 1X2 outcome frequencies by league (%), modelling period.

League	N	Home	Draw	Away
England (Premier League)	7,600	45.7	24.2	30.1
Spain (La Liga)	7,220	46.7	24.9	28.4
Italy (Serie A)	7,600	44.5	26.5	29.0
Germany (Bundesliga)	6,120	44.7	25.1	30.2
France (Ligue 1)	7,350	44.4	27.4	28.2
Portugal (Primeira Liga)	5,320	44.2	25.2	30.6
All core leagues	41,210	45.1	25.6	29.3

4.2.3 Shot-based features and their coverage

Shot counts and shots on target are available for the majority of the sample but are not uniformly distributed across competitions. In the five non-Portuguese core leagues fewer than 0.5% of modelling rows lack shot data; in the Portuguese Primeira Liga the corresponding figure is 54.2%, because the source does not record shot statistics for that competition before the 2017/18 season.

The gap falls entirely within the training period. Both the validation window used to fit the calibration maps and the test window used to evaluate them have complete coverage, so no calibration map is estimated from imputed data and no score reported in this chapter is computed from it. More importantly for the comparison at hand, all five calibration treatments are applied to an identical set of base-model outputs, so any weakening of the base learner attributable to incomplete covariates is shared by every treatment and cannot favour one over another. The analysis is repeated with shot-derived features removed in Section 4.6.2, where the base learners are correspondingly weaker and the pattern of results is unchanged in direction.

The final specification comprises 78 features derived from Elo ratings, rolling form over three, five and ten matches, rolling goal and shot statistics, rest days, fixture congestion, result streaks and Elo momentum, together with their home-minus-away differentials.

4.2.4 The bookmaker margin

Figure 4.1 reports the mean bookmaker overround by year for the single bookmaker whose prices are used throughout this study.

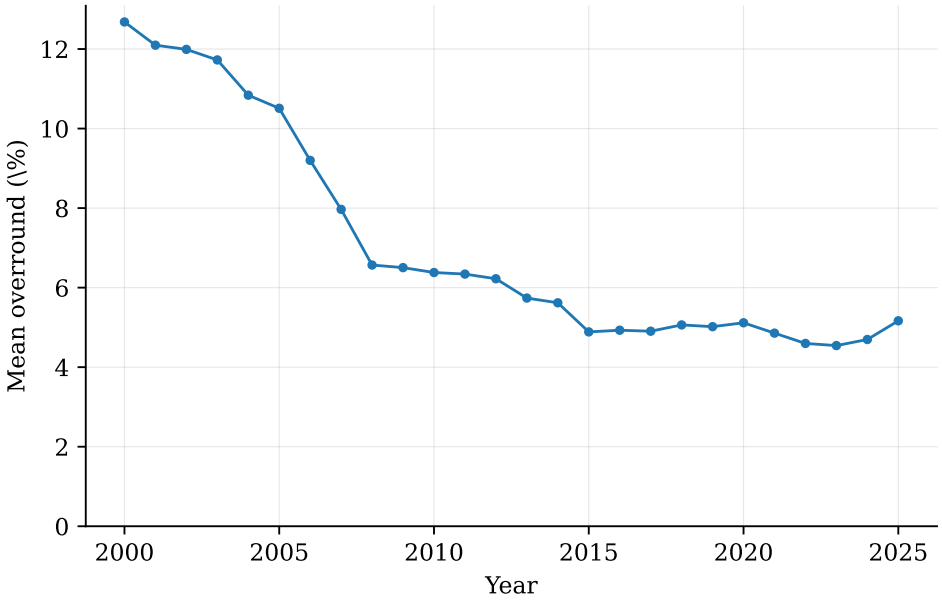


Figure 4.1: Mean bookmaker overround by year.

The single-bookmaker margin has compressed markedly over the sample period, from approximately 12.7% in 2000 to approximately 5.2% in recent seasons, with the sharpest decline occurring between 2003 and 2008, consistent with the account of increasing competition among online operators discussed in Chapter 2.

The mean overround across the sample is 1.068, and this figure is the one that conditions the interpretation of any economic reading of the results. The dataset also records the best price available across a panel of bookmakers, which implies a substantially smaller margin, but those prices are recorded only from 2005 and intermittently thereafter, and obtaining them consistently would require accounts with numerous operators that remain unrestricted, which is not generally the case for consistently winning bettors (Hegarty & Whelan, 2025). The conservative single-book prices are therefore used throughout.

4.2.5 Calibration of the market forecast

Because the de-vigged market forecast serves as the benchmark against which model probabilities are judged, its own calibration constitutes an empirical question rather than an assumption. Figure 4.2 presents the reliability diagram for market probabilities obtained by multiplicative de-vigging over the validation and test periods, where points lying on the diagonal indicate agreement between predicted probabilities and observed frequencies. All three outcome curves track the diagonal closely across the range of predicted probabilities in which observations are numerous. The market forecast is therefore well calibrated in the sense defined in Chapter 2.

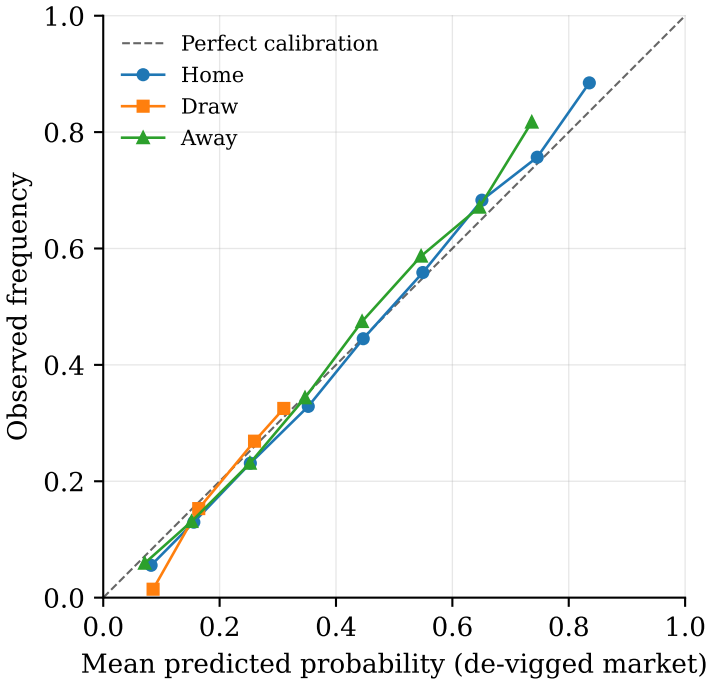


Figure 4.2: Reliability diagram for de-vigged market probabilities under multiplicative normalisation, validation and test periods combined.

This establishes a demanding standard and frames the question addressed in the remainder of this chapter. The relevant issue is not whether calibrated probabilities for football outcomes are attainable in principle - the market demonstrates that they are - but whether post-hoc calibration can bring machine-learning forecasts closer to that standard, and whether the choice among calibration methods materially affects the outcome.

4.3 Uncalibrated base-model performance

Two baselines are included alongside the machine-learning models. The first predicts the training-period class frequencies for every match and represents the floor below which no forecast should fall. The second is a multinomial logistic regression on the pre-match Elo difference alone, following Hvattum and Arntzen (2010), and represents the performance obtainable from team strength without further feature engineering.

Table 4.4: Out-of-sample probabilistic performance on the test period by base model and calibration method. **Bold** indicates the lowest (best) value in each column across the full table.

Model	Calibration	Log loss	Brier	RPS	ECE	Acc. (%)
Base rate	Raw	1.0788	0.6534	0.2314	0.0283	42.7
	Platt (OvR)	1.0770	0.6521	0.2307	0.0158	42.7
	Vector scaling	1.0770	0.6521	0.2307	0.0158	42.7
	Dirichlet	1.0770	0.6521	0.2307	0.0158	42.7
	Isotonic (OvR)	1.0770	0.6521	0.2307	0.0158	42.7
Elo logit	Raw	0.9747	0.5809	0.1961	0.0153	52.9
	Platt (OvR)	0.9730	0.5795	0.1954	0.0122	53.4
	Vector scaling	0.9727	0.5791	0.1953	0.0107	53.4
	Dirichlet	0.9728	0.5793	0.1954	**0.0089**	53.4
	Isotonic (OvR)	0.9916	0.5804	0.1958	0.0194	53.4
Market (de-vigged)	-	**0.9540**	**0.5663**	**0.1896**	0.0220	54.7
Random Forest	Raw	0.9711	0.5780	0.1950	0.0193	53.2
	Platt (OvR)	0.9693	0.5767	0.1943	0.0174	53.3
	Vector scaling	0.9691	0.5766	0.1943	0.0192	53.3
	Dirichlet	0.9700	0.5772	0.1945	0.0161	53.2
	Isotonic (OvR)	0.9892	0.5781	0.1948	0.0244	53.6
XGBoost	Raw	0.9719	0.5785	0.1951	0.0153	53.3
	Platt (OvR)	0.9694	0.5768	0.1943	0.0164	53.6
	Vector scaling	0.9692	0.5767	0.1943	0.0197	53.6
	Dirichlet	0.9694	0.5768	0.1943	0.0223	53.7
	Isotonic (OvR)	1.0021	0.5793	0.1950	0.0181	53.4

The uncalibrated results appear in the rows labelled "Raw" in Table 4.4. The base-rate forecast attains a test log loss of 1.0788, the Elo-only logistic regression 0.9747, the tuned Random Forest 0.9711 and the tuned XGBoost 0.9719.

The improvement obtained by moving from a single-covariate Elo model to a gradient-boosted ensemble over 78 engineered features is small: a reduction in log loss of approximately 0.3%. This is not a defect of the pipeline but a substantive finding, consistent with the observation that reported gains in this literature are frequently attributable to feature availability rather than algorithmic sophistication (Bunker & Susnjak, 2022). The greater part of the predictable variation in football outcomes is captured by relative team strength; form, schedule congestion and momentum contribute comparatively little once strength is controlled for. The two base learners also differ markedly in their propensity to overfit. The Random Forest attains a training log loss of 0.8866 against 0.9711 on test, a gap of 0.085, whereas XGBoost attains 0.9447 against 0.9719, a gap of 0.027. The two models generalise almost identically despite this difference. This is relevant to what follows, since models fitting their training data closely are often supposed to be overconfident out of sample and might be expected to benefit more from calibration.

One further number in Table 4.4 is worth extracting: the accuracy of the base-rate forecast, 42.7%. This forecast assigns the single most frequent outcome in the training period, a home win, to every match in the test period, and its accuracy is therefore simply the proportion of matches in the test period that were, in fact, home wins.

4.4 The effect of calibration

4.4.1 Experimental design

Each calibration map was fitted on the validation period of 4,264 matches and applied to the held-out test period of 4,115 matches. The test period was used exclusively for final evaluation and at no point informed the fitting of a model or a calibrator.

The five treatments of Table 4.1 differ in how they treat the three-class structure. Platt scaling and isotonic regression are binary techniques; their application here fits an independent map for each outcome and rescales the three outputs to sum to unity, a corrective applied after the fact rather than a property of the methods themselves. Dirichlet calibration is natively multiclass: following Kull et al. (2019), it is implemented as a multinomial logistic regression on the logarithm of the raw probability vector,

$$\mathbf{p}_{\text{cal}} = \text{softmax}(W \log \mathbf{p}_{\text{raw}} + \mathbf{b}) \,, \tag{4.1}$$

so that each calibrated probability depends upon all three raw probabilities and the output is a valid distribution by construction.

Comparing Platt scaling directly against full Dirichlet calibration would, however, confound two distinct properties. Full Dirichlet fits a 3×3 matrix and three intercepts, twelve parameters against Platt's six. A defeat for Dirichlet could therefore reflect either the irrelevance of native multiclass structure or the cost of the additional parameters. To separate these, a fourth treatment is introduced: vector scaling, which is Equation 4.1 with W constrained to be diagonal. It retains the softmax formulation and the guaranteed-valid output while fitting three scale parameters and three intercepts - exactly matching Platt's parameter count. The contrast between Platt scaling and vector scaling therefore isolates the effect of native multiclass structure at fixed complexity, and the contrast between vector scaling and full Dirichlet isolates the effect of complexity at fixed structure.

Table 4.4 reported out-of-sample performance for all treatments. Table 4.5 reports pairwise significance for the XGBoost base learner, and Table 4.6 the corresponding comparisons for the Random Forest, including the contrast between vector scaling and full Dirichlet calibration on which the complexity argument of Section 4.5 turns.

Table 4.5: Pairwise differences in mean per-match logarithmic score between calibration treatments, xgboost base learner, test period.

Comparison	Mean difference	95% CI	*p*-value
Raw vs. Platt	+0.0025*	[+0.0011, +0.0037]	0.001
Raw vs. Vector Scaling	+0.0027*	[+0.0012, +0.0040]	0.001
Raw vs. Isotonic	-0.0302*	[-0.0611, -0.0050]	0.012
Raw vs. Dirichlet	+0.0025*	[+0.0008, +0.0041]	0.004
Platt vs. Vector Scaling	+0.0002	[-0.0000, +0.0005]	0.099
Platt vs. Isotonic	-0.0327*	[-0.0638, -0.0075]	0.002
Platt vs. Dirichlet	+0.0000	[-0.0006, +0.0007]	1.000
Vector Scaling vs. Isotonic	-0.0329*	[-0.0641, -0.0077]	0.001
Vector Scaling vs. Dirichlet	-0.0002	[-0.0007, +0.0003]	0.423
Isotonic vs. Dirichlet	+0.0327*	[+0.0075, +0.0637]	0.001

Table 4.6: Pairwise differences in mean per-match logarithmic score between calibration treatments, Random Forest base learner, test period.

Comparison	Mean difference	95% CI	*p*
Raw vs. Platt	+0.0017*	[+0.0004, +0.0029]	0.009
Raw vs. Vector scaling	+0.0019*	[+0.0005, +0.0032]	0.009
Raw vs. Isotonic	-0.0182*	[-0.0392, -0.0004]	0.038
Raw vs. Dirichlet	+0.0011	[-0.0005, +0.0026]	0.181
Platt vs. Vector scaling	+0.0002	[-0.0000, +0.0004]	0.090
Platt vs. Isotonic	-0.0199*	[-0.0407, -0.0022]	0.003
Platt vs. Dirichlet	-0.0007	[-0.0014, +0.0001]	0.073
Vector scaling vs. Isotonic	-0.0201*	[-0.0408, -0.0024]	0.002
Vector scaling vs. Dirichlet	-0.0008*	[-0.0015, -0.0002]	0.012
Isotonic vs. Dirichlet	+0.0192*	[+0.0013, +0.0404]	0.015

4.4.2 Calibration produces small but statistically reliable gains

Post-hoc calibration improves probabilistic performance, but modestly. For XGBoost, vector scaling reduces log loss from 0.9719 to 0.9692, and Platt scaling and Dirichlet calibration each to 0.9694; Brier score falls from 0.5785 to 0.5767 and RPS from 0.1951 to 0.1943. For the Random Forest, vector scaling attains 0.9691 against a raw 0.9711, the best result recorded for any model-calibration combination in this study. The Elo baseline improves from 0.9747 to 0.9727.

These improvements are small in magnitude but not attributable to chance. The block bootstrap of Table 4.5, which resamples the 21 calendar months of the test period rather than individual matches and thereby respects the temporal dependence that a paired test would ignore, places the raw-versus-Platt difference for XGBoost at 0.0025 with a 95% interval of $[0.0011, 0.0037]$ and $p = 0.001$, and the raw-versus-vector-scaling difference at 0.0027, $[0.0012, 0.0040]$, $p = 0.001$. The corresponding raw-versus-Dirichlet difference is 0.0025, $[0.0008, 0.0041]$, $p = 0.004$.

For the Random Forest, Platt scaling and vector scaling both improve on the raw forecast at conventional significance, at 0.0017 and 0.0019 respectively with $p = 0.009$ in each case, but full Dirichlet calibration does not: the difference is 0.0011 with an interval of $[-0.0005, 0.0026]$ and $p = 0.181$. The statement that calibration reliably improves probabilistic performance therefore holds for the two six-parameter maps on both tree ensembles, and for the twelve-parameter map on XGBoost only. This is the first indication of the pattern developed in Section 4.5.

Figure 4.3 presents the same comparisons graphically. Intervals excluding zero denote a distinguishable difference; the contrasts involving isotonic regression lie far from zero, whereas those among the parametric treatments straddle it.

The limited magnitude of the gain indicates that the raw forecasts were already reasonably well calibrated. This is unsurprising for XGBoost, which optimises the multiclass logarithmic score directly and is therefore trained to produce honest probabilities rather than merely separable classes. It is more surprising for the Random Forest, whose considerably larger training-test gap suggested overconfidence; yet the Random Forest gains no more from calibration than XGBoost does. Fitting training data closely evidently does not, in this instance, translate into systematically miscalibrated out-of-sample probabilities.

A further observation supports the framing adopted throughout this dissertation. Accuracy moves scarcely at all across treatments: for XGBoost between 53.3% and 53.7%, for the Random Forest between 53.2% and 53.6%. Calibration is a monotone or near-monotone transformation that rarely alters which outcome is deemed most probable. The improvements recorded in the proper scoring rules are therefore improvements in probability quality that accuracy is structurally incapable of detecting.

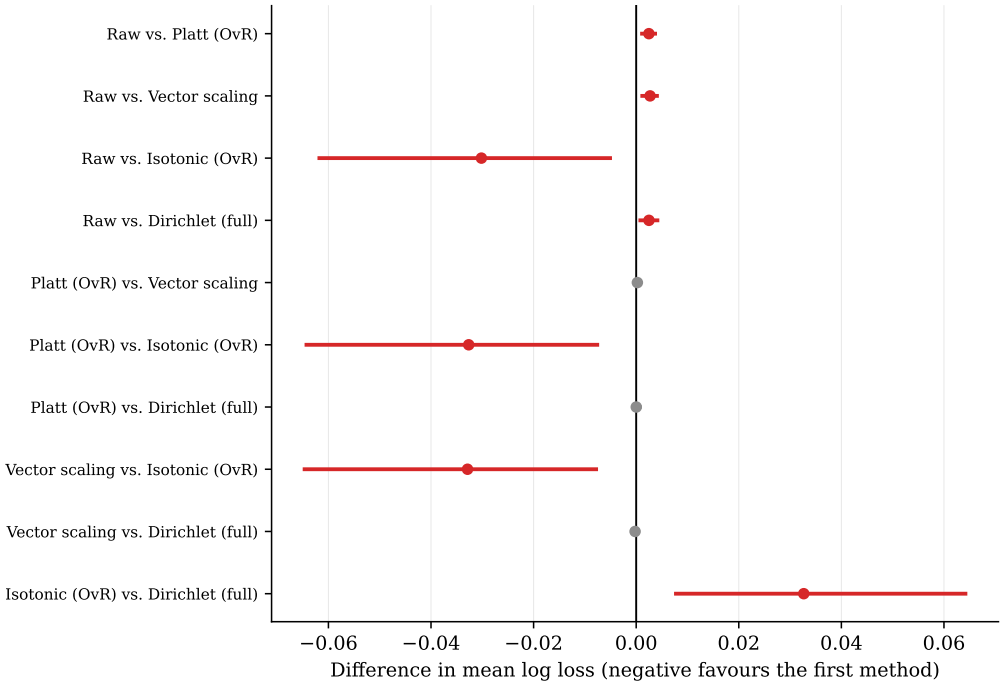


Figure 4.3: Pairwise differences in mean per-match logarithmic score with 95% block-bootstrap intervals, XGBoost base learner.

4.4.3 The parametric maps are consistently ordered but negligibly separated

The three parametric treatments are designed to separate two properties that a conventional comparison confounds. Platt scaling and vector scaling fit the same number of parameters but differ in structure, so the contrast between them isolates the effect of native multiclass formulation. Vector scaling and full Dirichlet calibration share their functional form, their objective, their penalty and their optimiser, differing only in whether the off-diagonal elements of W are free, so the contrast between them isolates the effect of complexity.

Vector scaling attains the lowest logarithmic score of the three for every base learner: 0.9692 against 0.9694 for both Platt scaling and full Dirichlet calibration under XGBoost; 0.9691 against 0.9693 and 0.9700 under the Random Forest; and 0.9727 against 0.9730 and 0.9728 for the Elo baseline. The spread across the three maps is 0.0002 for XGBoost, 0.0009 for the Random Forest and 0.0003 for the Elo baseline. Both pre-specified contrasts therefore point in the direction the design anticipated, on both tree ensembles: the natively multiclass map improves on the one-versus-rest map at fixed parameter count, and the six-parameter map improves on the twelve-parameter map at fixed structure.

The block bootstrap places almost all of this within the resolution of the test period. The structural contrast is 0.0002 in favour of vector scaling for both ensembles, at $[-0.0000, 0.0005]$, $p = 0.099$ for XGBoost and $[-0.0000, 0.0004]$, $p = 0.090$ for the Random Forest; neither

reaches conventional significance, and the two intervals are almost identical. The complexity contrast is 0.0002 in favour of vector scaling for XGBoost, at $[-0.0007, 0.0003]$, $p = 0.423$, and 0.0008 for the Random Forest, at $[-0.0015, -0.0002]$, $p = 0.012$. Only that last comparison excludes zero. Platt scaling and full Dirichlet calibration are indistinguishable for XGBoost (0.0000, $p = 1.000$) and marginal for the Random Forest (0.0007 in favour of Platt scaling, $p = 0.073$).

Consistency of sign across base learners is weaker evidence than the count of comparisons suggests. The three base learners are fitted to the same covariates and evaluated on the same 4,115 matches, and their raw probability vectors are strongly correlated; a calibration map that happens to suit one set of raw outputs will tend to suit the others, so four contrasts pointing the same way are not four independent confirmations. What the evidence supports is that the ordering is stable rather than arbitrary, and that its magnitude is between 0.0002 and 0.0008 in logarithmic score.

The largest separation among the parametric maps, 0.0008 for the Random Forest, is roughly a third of the gain that calibration itself secures over the raw forecast and around one twentieth of the damage inflicted by isotonic regression. A practitioner choosing among Platt scaling, vector scaling and full Dirichlet calibration is choosing among maps whose realised difference is smaller than the run-to-run variation reported for the smaller calibration windows of Section 4.5. The choice among them has no practical consequence, and the single significant contrast does not alter that: it establishes that the simpler of two nested maps is detectably better on one base learner, not that the difference matters.

It is worth being explicit about which hypothesis this does and does not support. The complexity account developed in Section 4.5 is strengthened, in that the more heavily parameterised map is behind on both ensembles and detectably so on one, and in that it is the only treatment failing to improve on the raw forecast at conventional significance (Table 4.6). The structural account is not established: the natively multiclass map leads on both ensembles by an identical 0.0002, which is a difference too small to attribute to the mechanism that motivated it.

4.4.4 Isotonic regression degrades performance substantially

The largest effect in Table 4.4 is negative. Isotonic regression does not merely fail to improve upon the raw forecasts; it degrades them, for every base learner, by a margin an order of magnitude larger than any method's improvement. For XGBoost, log loss rises from 0.9719 to 1.0021, leaving the calibrated forecast worse than the uncalibrated one and worse than the Elo-only baseline. The Random Forest deteriorates from 0.9711 to 0.9892, and the Elo logistic regression from 0.9747 to 0.9916. The block bootstrap confirms that these are not sampling artefacts: for XGBoost the raw-versus-isotonic difference is -0.0302,

$[-0.0611, -0.0050]$, $p = 0.012$.

The failure is confined almost entirely to the logarithmic score. Under the Brier score isotonic regression is approximately neutral: 0.5793 against a raw 0.5785 for XGBoost, and 0.5781 against 0.5780 for the Random Forest. The divergence between scoring rules is itself informative. The Brier score is bounded and penalises probabilistic error quadratically, whereas the logarithmic score is unbounded and diverges as an assigned probability approaches zero for an outcome that occurs. Isotonic regression fits a step function and can map entire regions of the probability space to values at or near zero; whenever the corresponding outcome materialises, the logarithmic penalty is severe. A method may accordingly appear acceptable under one proper scoring rule while performing poorly under another, which argues for reporting several.

The divergence extends further than the Brier score. Under RPS, isotonic regression is not merely neutral but marginally superior to the uncalibrated forecast, and for every base learner: 0.1950 against 0.1951 for XGBoost, 0.1948 against 0.1950 for the Random Forest, and 0.1958 against 0.1961 for the Elo baseline. Three strictly proper scoring rules therefore return three different verdicts on the same calibration map. The logarithmic score judges it catastrophic, the Brier score approximately neutral, and RPS marginally beneficial.

The explanation lies in what each rule rewards, as RPS evaluates the cumulative distribution rather than the probability assigned to the realised outcome, and is consequently insensitive to the near-zero probabilities that drive the logarithmic penalty: a forecast can place negligible mass on the true outcome and still accumulate little RPS penalty provided the mass it does place is adjacent to it. Isotonic regression, which sharpens the ordering of predictions while occasionally assigning extreme values, is flattered by this and punished by the logarithmic score.

Proper scoring rules can rank methods differently in principle and on the same data, the same predictions and the same calibration map, and the choice among three standard and strictly proper rules determines whether a method is judged the worst available or marginally the best. Reporting a single rule is not a matter of presentational completeness but of whether the reported conclusion is the correct one.

This property has an implementation consequence that must be disclosed. Because isotonic regression produces probabilities at or near zero, the numerical floor applied before taking logarithms materially affects the reported score. Recomputing XGBoost's isotonic log loss under floors of 10^{-6}, 10^{-9} and 10^{-12} yields 0.9854, 0.9938 and 1.0021 respectively, the value saturating at 10^{-12} because that is the floor applied within the calibrator itself. Roughly half of the headline degradation is therefore attributable to the flooring convention rather than to the calibration map alone. The finding is nonetheless robust in sign and in significance: under the most permissive floor considered, isotonic regression still degrades log loss

by 0.0135 relative to the uncalibrated forecast, and remains the worst-performing treatment by a wide margin. No comparable sensitivity affects the other methods, whose outputs are bounded away from zero by construction. The reported figures use a floor of 10^{-12} throughout.

The deterioration is also largest for the sharpest base learner. XGBoost, whose raw probability distribution is the most concentrated, suffers the greatest degradation. This points towards an explanation in terms of calibration-set size, developed in the next section.

4.5 Calibration-map complexity and calibration-set size

The results of Section 4.4 raise a question that the test-period comparison alone cannot answer. If the three parametric maps are indistinguishable from one another while the non-parametric map fails badly, is the operative variable the functional form of the map or the quantity of data available to fit it? Figure 4.4 addresses this directly by refitting each calibrator on subsamples of the validation period and evaluating on the unchanged test period, where each point averages twenty random subsamples of the validation period, and bars denote one standard deviation - the dashed line is the uncalibrated forecast:

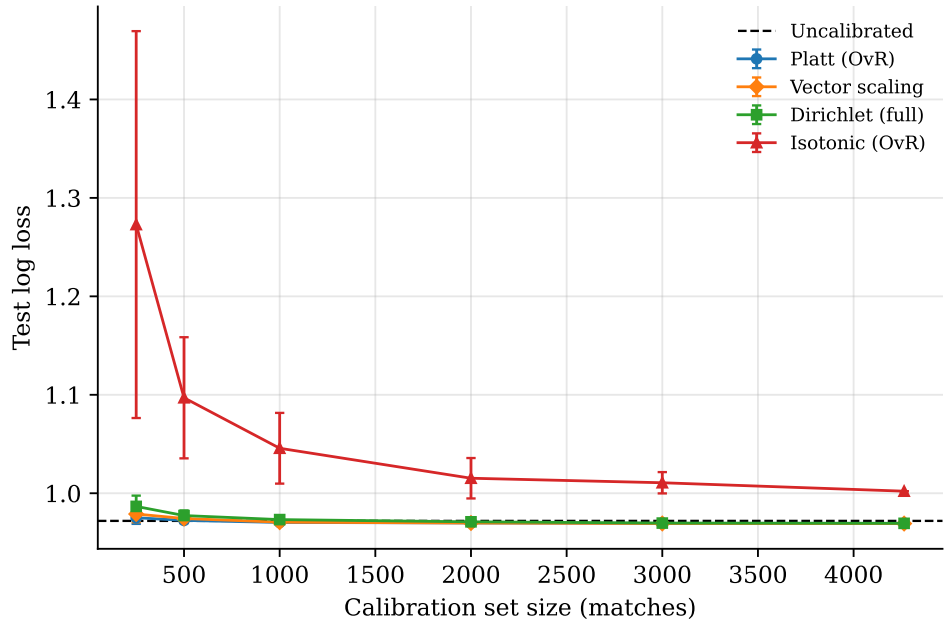


Figure 4.4: Test-period logarithmic score as a function of calibration-set size, XGBoost base learner.

At the smallest calibration set of 250 matches, every calibration method is worse than no calibration at all. The uncalibrated XGBoost forecast attains 0.9719; Platt scaling attains 0.9751, vector scaling 0.9789, full Dirichlet 0.9866 and isotonic regression 1.2729. The same pattern holds for the Random Forest and for the Elo baseline.

At this sample size the parametric maps are ordered by complexity. The two six-parameter maps are separated from one another by 0.0038, whereas each is separated from the twelve-parameter map by more: 0.0077 for vector scaling and 0.0115 for Platt scaling. Doubling the parameter count therefore costs more at $n = 250$ than changing the structure of the map does, which is the ordering the complexity account predicts and which the full-window comparison of Section 4.4.3 is too coarse to reveal. The claim should not be pushed further than that: three points do not establish a functional relationship between parameter count and required sample size, and the two six-parameter maps do differ from one another, so parameter count is not the only variable operating.

The separation from the non-parametric map is of a different order - isotonic regression attains 1.2729 at $n = 250$, with a variability across subsamples far above that of any parametric map. It remains above the uncalibrated benchmark at every sample size examined, reaching 1.0021 with the full window of 4,264 matches and showing no sign of having converged. The parametric maps, by contrast, all cross below the uncalibrated benchmark and then converge on one another: Platt scaling and vector scaling cross between 500 and 1,000 matches, full Dirichlet between 1,000 and 2,000, and by $n = 3{,}000$ the three lie within 0.0004 of each other. The crossing point itself moves with complexity, which is the ordering of the previous paragraph observed from a different direction.

The pattern replicates across all three base learners, with the crossover point and the magnitude of the isotonic penalty varying little between them. This rules out an explanation specific to any one model's probability distribution. Table 4.7 reports the values underlying Figure 4.4.

Table 4.7: Test-period logarithmic score by calibration-set size, XGBoost base learner.

Method	Calibration set size					
	250	500	1,000	2,000	3,000	4,264
Platt (OvR)	0.9751	0.9724	0.9703	0.9698	0.9694	0.9694
Vector scaling	0.9789	0.9744	0.9706	0.9698	0.9693	0.9692
Dirichlet (full)	0.9866	0.9773	0.9732	0.9710	0.9697	0.9694
Isotonic (OvR)	1.2729	1.0970	1.0457	1.0152	1.0107	1.0021

The subsamples above are drawn at random from the validation period, which isolates the effect of size from that of recency but is not the situation a practitioner faces: a calibration window of 250 matches in deployment consists of 250 consecutive matches, which are less internally varied than 250 drawn at random from two seasons. Repeating the exercise with contiguous windows ending at the validation boundary gives uniformly worse results at small sizes - 0.9931 for Platt scaling at $n = 250$ against 0.9751 under random sampling - and the parametric methods do not cross below the uncalibrated benchmark until approxi-

mately 2,000 matches. The random-sampling results therefore understate the difficulty, and the operationally relevant threshold is higher than the headline curve suggests.

Two conclusions follow. The first is that calibration-map complexity orders the methods at every sample size examined, but that the practical weight of the ordering collapses as the window grows. At 250 matches the three parametric maps are separated by 0.0115 and ranked by parameter count; at 4,264 they are separated by 0.0002 and the ranking, though stable, is of no consequence. The sharp distinction is between parametric and non-parametric calibration, because only there does the ordering remain large at the full window. Complexity is the right variable; the calibration set available here is simply large enough to absorb the difference between six parameters and twelve, and nowhere near large enough to absorb an unbounded map.

The second is that the constraint is intrinsic to the problem rather than incidental to this design. A calibration window in a temporally ordered pipeline must be long enough to support the fitted map and recent enough to remain representative of the period on which it will be used. Two seasons across six major leagues yields roughly four thousand matches. Extending the window draws on progressively less representative seasons; contracting it moves the analysis into the region where the contiguous-window results show calibration to be actively harmful. The window used here is close to the largest that temporal validity permits, and a non-parametric map is still unusable within it.

4.6 Robustness

Several of the principal findings of this chapter are null or negative, and are correspondingly vulnerable to the objection that they reflect a particular analytical choice rather than a substantive effect. Two robustness analyses address that objection directly: a sweep of the regularisation applied to the Dirichlet map, and a re-run of the entire pipeline with the shot-derived features removed. A third subsection decomposes calibration error by outcome class; that is a diagnostic rather than a robustness check, but it is reported here because it revises the interpretation of the results above.

4.6.1 Sensitivity to regularisation

The comparison between full Dirichlet calibration and its diagonally constrained counterpart depends on the penalty applied to W. Kull et al. (2019) recommend penalising the off-diagonal and intercept terms specifically, on the grounds that the full matrix parameterisation is prone to overfitting; the implementation used here applies an L_2 penalty shrinking W toward the identity map and $\mathbf{b}$ toward zero, identically for both treatments.

Table 4.8 reports test performance across five orders of magnitude of penalty strength

for the XGBoost base learner. The result is insensitivity rather than a trade-off: log loss is 0.9694 at every value of λ from 10^{-4} to 1, with Brier score and RPS likewise unchanged to four decimal places, and moves only at $\lambda = 10$, where it reaches 0.9697. The map fitted to 4,264 matches is evidently determined by the data rather than by the penalty across the whole range in which the penalty is plausible. The value used throughout, $\lambda = 10^{-2}$, could be varied by four orders of magnitude in either direction without altering any reported figure, so the Dirichlet results of Section 4.4.3 are not attributable to the choice of penalty strength.

Table 4.8: Sensitivity of Dirichlet calibration to the strength of the L_2 penalty, XGBoost base learner, test period.

λ	Log loss	Brier	RPS
0.0001	0.9694	0.5768	0.1943
0.001	0.9694	0.5768	0.1943
0.01	0.9694	0.5768	0.1943
0.1	0.9694	0.5768	0.1943
1	0.9694	0.5768	0.1943
10	0.9697	0.5769	0.1944

Insensitivity to the strength of an L_2 penalty does not establish insensitivity to its structure, and the off-diagonal scheme of Kull et al. (2019) is a different constraint rather than a differently weighted one. The vector scaling treatment bears on this. Constraining the off-diagonal terms to zero is the limiting case of penalising them, and it changes the score by 0.0002 for XGBoost and 0.0008 for the Random Forest. The limiting case therefore produces a small but, on one base learner, detectable improvement, which suggests that an intermediate constraint of the kind Kull et al. (2019) recommend would fall somewhere between the two and would not alter the ordering. It has not been tested directly.

4.6.2 Sensitivity to the shot-feature specification

Section 4.2.3 documents that shot statistics are retained despite uneven coverage, on the grounds that the gap falls entirely within the training period and that all calibration treatments operate on identical base-model outputs. The entire pipeline was repeated with shot-derived features removed, reducing the specification from 78 features to 51, in order to test whether the conclusions depend on that choice.

Removing them weakens the base learners, as expected: the Random Forest test log loss rises from 0.9711 to 0.9748 and XGBoost from 0.9719 to 0.9756. The deterioration of roughly 0.004 exceeds the entire effect of calibration, which is worth recording in its own right, since it indicates that the feature specification matters more to raw forecast quality than the calibration step that follows it.

Isotonic regression remains the worst treatment for every base learner, and the parametric maps remain separated by margins of the kind reported in Section 4.4.3. What changes is the magnitude of the isotonic penalty, which falls from 0.0302 to 0.0052 for XGBoost and from 0.0182 to 0.0165 for the Random Forest, and which ceases to be statistically distinguishable from the uncalibrated forecast in either case ($p = 0.482$ and $p = 0.060$ respectively).

Section 4.4.4 attributes the isotonic failure in part to the sharpness of the underlying forecast, on the grounds that a sharper base learner produces more extreme probabilities for a step function to map to zero. Removing 27 informative features produces a blunter base learner, and for XGBoost - the sharper of the two ensembles, and the one whose raw distribution is most concentrated - the isotonic penalty falls roughly sixfold. For the Random Forest, whose forecasts were less concentrated to begin with, it falls by less than a tenth. The specification comparison therefore functions as an unintended replication of the proposed mechanism, and the asymmetry between the two ensembles is itself part of the evidence: the damage inflicted by an over-flexible calibration map scales with the sharpness of the forecast it is applied to.

The reported specification retains shot features on the reasoning given in Section 4.2.3.

4.6.3 Class-wise calibration and the draw

Aggregate measures can conceal outcome-specific miscalibration, and the draw class has particular salience given its documented difficulty. Table 4.9 decomposes calibration error by outcome.

The class-wise decomposition anticipating that the draw would emerge as the worst-calibrated outcome, consistent with the difficulty the literature reports, found the opposite as true. The draw is the best calibrated of the three outcomes for every real forecaster in the study: for both tree ensembles and the Elo baseline, under every calibration treatment. For raw XGBoost the class-wise errors are 0.0318 for away wins, 0.0159 for draws and 0.0368 for home wins; for the Random Forest, 0.0321, 0.0085 and 0.0356. The exception is the degenerate base-rate forecast, where the calibrated variants place the away class marginally ahead of the draw, at 0.0073 against 0.0085; a forecast that issues the same three numbers for every match has no class-wise structure to speak of. The class that the literature identifies as hardest to predict is, on this measure, the class whose probabilities are most trustworthy.

Table 4.9: Class-wise expected calibration error on the test period.

Model	Calibration	ECE (Away)	ECE (Draw)	ECE (Home)
Base rate	Dirichlet	0.0073	0.0085	0.0158
	Isotonic (OvR)	0.0073	0.0085	0.0158
	Platt (OvR)	0.0073	0.0085	0.0158
	Raw	0.0285	0.0002	0.0283
	Vector scaling	0.0073	0.0085	0.0158
Elo logit	Dirichlet	0.0242	0.0094	0.0225
	Isotonic (OvR)	0.0207	0.0128	0.0260
	Platt (OvR)	0.0209	0.0100	0.0220
	Raw	0.0263	0.0060	0.0309
	Vector scaling	0.0207	0.0095	0.0221
Random Forest	Dirichlet	0.0257	0.0101	0.0278
	Isotonic (OvR)	0.0251	0.0106	0.0286
	Platt (OvR)	0.0267	0.0108	0.0242
	Raw	0.0321	0.0085	0.0356
	Vector scaling	0.0263	0.0109	0.0244
XGBoost	Dirichlet	0.0282	0.0108	0.0235
	Isotonic (OvR)	0.0313	0.0183	0.0252
	Platt (OvR)	0.0251	0.0148	0.0232
	Raw	0.0318	0.0159	0.0368
	Vector scaling	0.0269	0.0147	0.0227

Draw probabilities in this sample occupy a narrow band around the base rate of roughly one quarter, because the covariates carry little information discriminating draws from other outcomes. Predictions that barely vary leave little room for predicted and observed frequencies to diverge, so a low calibration error follows almost mechanically from the narrowness of the band rather than from any accuracy in placing the probabilities within it. Home and away probabilities range far more widely, offer correspondingly more scope for miscalibration, and duly exhibit it. Chapter 5 takes up what this implies for the received account of the draw.

Figure 4.5 makes this visible directly and Table 4.10 quantifies it. The dispersion of predicted draw probabilities is approximately one third that of the two win outcomes: a standard deviation of 0.058 against 0.165 for away wins and 0.185 for home wins, and a 5th-to-95th percentile range of 0.191 against 0.556 and 0.633 respectively. Predicted draw probabilities are effectively confined to the interval $[0.12, 0.31]$, whereas the win outcomes range across the greater part of the unit interval. The low measured calibration error on the draw class is therefore a consequence of limited dispersion rather than of superior probability estimation: a forecast that varies little around the base rate has correspondingly little opportunity to depart from it.

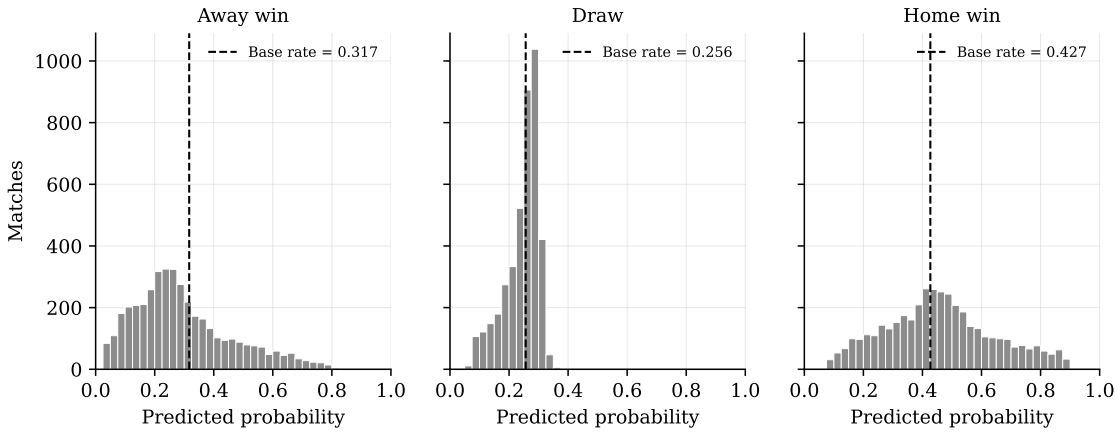


Figure 4.5: Distribution of predicted probabilities by outcome class, uncalibrated XGBoost, test period.

Table 4.10: Dispersion of predicted probabilities by outcome class, XGBoost base learner, test period.

Outcome	Base rate	Mean	SD	5th-95th pct.	90% range
Away win	0.317	0.296	0.165	[0.078, 0.634]	0.556
Draw	0.256	0.244	0.058	[0.120, 0.311]	0.191
Home win	0.427	0.460	0.185	[0.164, 0.797]	0.633

Figure 4.6 shows the corresponding reliability diagrams for each calibration treatment, permitting class-wise behaviour to be inspected alongside the aggregate measures.

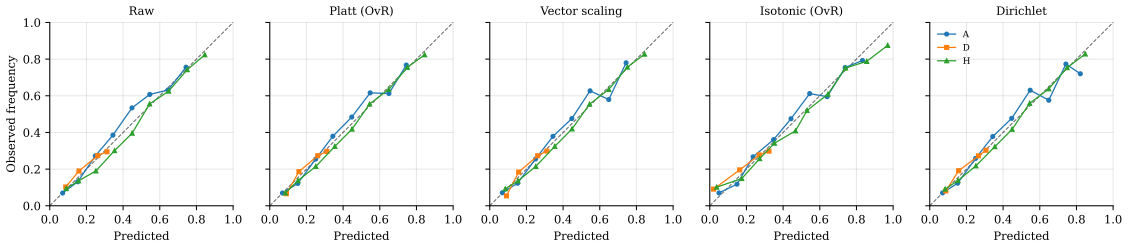


Figure 4.6: Reliability diagrams by calibration treatment, XGBoost base learner, test period. Points on the diagonal indicate agreement between predicted probabilities and observed frequencies.

This has a bearing on the calibration comparison as a whole. If the draw class contributes little variation in predicted probabilities, it also contributes little to the differences between calibration methods - which helps explain both why the natively multiclass formulation secures only a marginal advantage and why the vector scaling parameters adjust the draw dimension almost exclusively.

4.6.4 Expected calibration error does not track probabilistic performance

The previously shown Table 4.4 additionally reports expected calibration error, and the pattern it exhibits runs counter to every other measure in the study. For XGBoost the raw forecast attains the lowest ECE of the five treatments, at 0.0153, while simultaneously attaining the highest log loss, at 0.9719. Dirichlet calibration inverts the ordering, attaining the highest ECE at 0.0223 while sitting within 0.0002 of the best logarithmic score in the group. The ranking induced by ECE therefore contradicts the ranking induced by every proper scoring rule and by the significance tests of Table 4.5.

Figure 4.7 plots the two measures against one another for every pipeline in the study. A measure suitable for selecting among forecasters would produce an upward trend, better logarithmic scores accompanying better calibration error; no such trend is present.

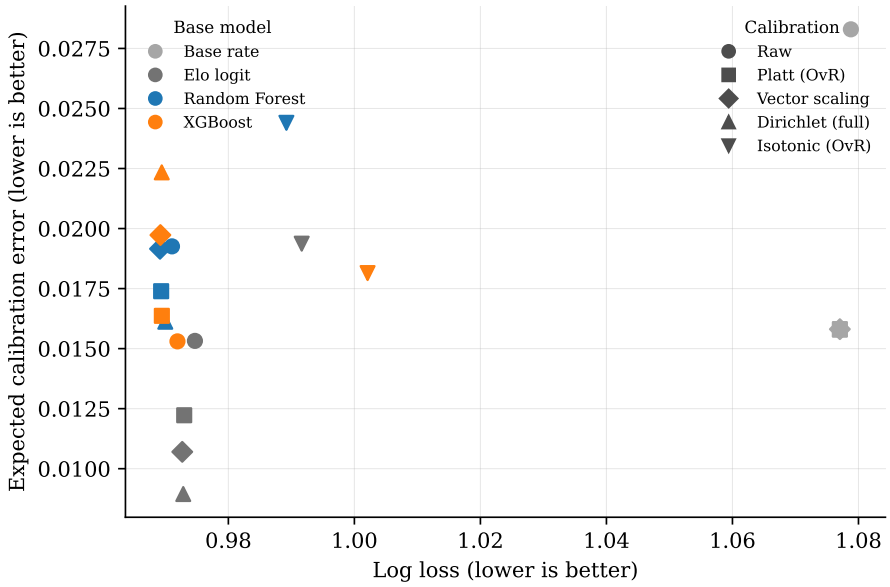


Figure 4.7: Expected calibration error against logarithmic score for every model-calibration pipeline, test period.

ECE as computed here summarises the calibration of the predicted-class confidence alone and discards information about the remainder of the probability vector, whereas log loss, Brier score and RPS evaluate the vector in full. More fundamentally, ECE is not a proper scoring rule: a forecaster can attain a low ECE by issuing probabilities close to the base rate - being reliably unconfident - without conveying useful information. The base-rate forecast illustrates this directly, attaining a calibrated ECE of 0.0158, superior to that of several considerably more informative forecasts, while remaining the worst forecast in the table by every proper scoring rule. Calibration is desirable only conditional on sharpness, and a measure that rewards the former in isolation will mislead. Compounded by the instability with respect

to binning noted in Chapter 2, ECE is reported here as a diagnostic rather than as a criterion for ranking methods.

4.7 Comparison with the market

The de-vigged market forecast attains a test log loss of 0.9540, a Brier score of 0.5663, an RPS of 0.1896 and an accuracy of 54.7%. On every proper scoring rule it outperforms every model-calibration combination examined. The margin is not negligible: the gap of 0.0151 in log loss between the market and the best calibrated model - the Random Forest under vector scaling, at 0.9691 - is approximately six times the gain that the best calibration method secured over its raw counterpart.

This is consistent with the weight of the literature and is not a defect of the framework. Bookmaker prices aggregate information unavailable in the present feature set - team news, confirmed lineups, injuries and the aggregate judgement of a liquid market - and are set by operators with a direct financial incentive to price accurately. That a model constructed from publicly available match results and ratings does not surpass such a forecast is the expected outcome rather than a failure of implementation.

Calibration narrows the gap without closing it. Raw XGBoost trails the market by 0.0179 in log loss; after vector scaling the deficit falls to 0.0152. The improvement is real but roughly a seventh of the deficit it does not address, and it leaves the ordering unchanged.

The market's superiority is also confined to the proper scoring rules. Its expected calibration error, at 0.0220, exceeds that of most calibrated models. Interpreted in light of Section 4.6.3, this is not evidence that the market is poorly calibrated but a further illustration that ECE penalises sharpness.

The market issues considerably more confident forecasts than the models do - its accuracy of 54.7% against a model range of 53.2 to 53.7% reflects genuinely superior discrimination - and a sharper forecast has more scope to accumulate measured calibration error even when it is, in the sense that matters for decision-making, the better forecast. The market is simultaneously the sharpest and the highest-scoring forecaster in the table, which is precisely the combination that Gneiting and Raftery (2007) identify as the objective: maximal sharpness subject to calibration.

Chapter 5

Discussion and Conclusion

5.1 Summary of the argument

5.1.1 Summary of the argument

This dissertation asked whether post-hoc probability calibration improves machine-learning forecasts of football 1X2 outcomes, and whether the choice among calibration methods materially affects the result. The answer to the first question is a qualified yes: calibration reduces the logarithmic score of a tuned XGBoost forecast by approximately 0.3% on average, but this reduction is small enough that it does not reach statistical significance under a block bootstrap for every pairwise comparison, and only one contrast - vector scaling against full Dirichlet calibration, for one base learner - excludes zero.

What was more interesting was the size of the gap between the best and worst calibration methods: the best treatments improve on the uncalibrated forecast by 0.0027 in log loss, while the worst degrades it by 0.0302, an order of magnitude larger than the average effect of calibrating at all. The strongest conclusion is therefore not that calibration reliably helps, but that the choice of calibration method carries more consequence, in either direction, than the decision to calibrate in the first place.

The remainder of this chapter interprets the principal findings, relates them to the literature reviewed in Chapter 2, acknowledges the limitations of the design, and identifies what follows.

5.2 Parametric versus non-parametric calibration

Methods are conventionally distinguished by their assumptions, parametric against non-parametric and binary against natively multiclass, and the theoretical case for the natively multiclass formulation is straightforward: 1X2 forecasting is a genuine three-class problem,

and adapting a binary method by renormalisation repairs an incoherence rather than avoiding it. The evidence here does not support that distinction being consequential, and does support a different one.

Among the parametric maps, the ordering is stable and the magnitude is negligible. Vector scaling attains the lowest logarithmic score for all three base learners, ahead of Platt scaling by 0.0002 in each case and ahead of full Dirichlet calibration by between 0.0002 and 0.0008. Both contrasts therefore point in the direction the design anticipated, which was not the expectation with which the analysis was run. Of the four comparisons only one excludes zero under the block bootstrap: vector scaling against full Dirichlet calibration for the Random Forest, at 0.0008 with $p = 0.012$. The others return $p = 0.099$, $p = 0.090$ and $p = 0.423$.

Two things follow, and they pull in opposite directions. Consistency of sign is evidence of something, but weaker evidence than four comparisons would suggest, because the base learners share their covariates, their calibration window and their test period, so their raw probability vectors are correlated and the four contrasts are not four independent trials. And the largest separation observed, 0.0008 in logarithmic score, is a third of what calibration itself achieves over the raw forecast and a twentieth of what isotonic regression destroys. The honest statement is that the ordering among parametric maps is real and stable, and that acting on it would be pointless.

The contrast that does carry weight is parametric against non-parametric. Isotonic regression degrades every base learner it is applied to, by roughly ten times the amount any parametric method improves on the uncalibrated forecast, and the learning curve shows it remains above the uncalibrated benchmark at every calibration-set size available, with a variability across subsamples far exceeding that of the parametric maps.

A calibration map is itself a statistical model fitted to a sample, subject to the same bias-variance considerations as any other, and the relevant question when selecting one is whether its flexibility is supported by the data available to fit it. That question has real force within the parametric family, but only at small samples. At 250 calibration matches the three parametric maps are separated by 0.0115 and ranked by parameter count, with the twelve-parameter map clearly worst; at the full window of 4,264 they are separated by 0.0002 and the same ranking is of no consequence. Four thousand observations are enough to absorb the difference between six parameters and twelve, and nowhere near enough to absorb an unbounded map. The variable is the same throughout; what changes is whether the calibration set is large enough to render it immaterial.

What makes this consequential for football forecasting specifically is that the constraint cannot be relieved by collecting more data. Section 4.5 sets out why: the two requirements a calibration window must satisfy, sufficient size and sufficient recency, are in direct tension, and six major European leagues over two seasons is roughly where the tension balances. Neither

direction offers an escape. A longer window recruits seasons progressively less like the one being forecast, and a shorter one moves the analysis into the region where the contiguous-window results show calibration doing active harm.

5.3 The failure of isotonic regression

Isotonic regression is widely recommended as a calibration method, and is the more flexible of the two options offered by default in scikit-learn's calibration interface. In this application it degrades every base learner it is applied to, leaving the calibrated XGBoost forecast worse than the uncalibrated one and worse than a single-covariate Elo baseline.

Two mechanisms combined to produce this. The first is the one discussed above: with 4,264 calibration matches distributed across three outcomes, the effective sample supporting each fitted step function is modest, and the map tracks noise that does not persist into the test period. The learning curve confirms this, showing isotonic regression far from convergence even at the largest window available and never crossing below the uncalibrated benchmark.

The second is specific to the interaction between isotonic regression and the logarithmic score. A step function can map an entire region of the probability space to zero, and the logarithmic score diverges when an outcome assigned zero probability occurs. This produces a large penalty concentrated in a small number of matches. The Brier score, being bounded, registers almost nothing: isotonic regression is approximately neutral under Brier while being catastrophic under log loss.

5.4 The multiclass hypothesis

The inclusion of Dirichlet calibration was motivated by the argument that 1X2 forecasting is a genuine three-class problem, and that binary methods adapted through renormalisation treat a constrained three-dimensional object as three unconstrained one-dimensional ones, repairing the resulting incoherence only after the fact. The natively multiclass formulation avoids this by construction.

The evidence is directionally consistent with the hypothesis and far too weak to support it. Vector scaling, the natively multiclass map fitted with the same parameter count as Platt scaling, attains the lower logarithmic score for both tree ensembles and for the Elo baseline, but the margin is 0.0002 in every case and neither ensemble comparison excludes zero ($p = 0.099$ and $p = 0.090$). A difference that reproduces at the same magnitude on three base learners is unlikely to be pure noise. A difference of 0.0002 in logarithmic score is also too small to attribute to the mechanism that motivated the hypothesis, or to anything else in particular. Full Dirichlet calibration, which shares the natively multiclass form but doubles

the parameters, sits behind vector scaling on both ensembles, so within the multiclass family the structural advantage does not survive the addition of parameters.

The objection to renormalisation concerns coherence, and renormalisation does restore coherence. The one-versus-rest maps arrive at valid probabilities by a less principled route. Whether the route matters is an empirical question, and at this sample size and on this problem the answer is that it matters by 0.0002.

The second point is specific to the geometry of football outcomes and connects to the analysis of the draw class. Native multiclass structure permits each calibrated probability to depend on all three raw probabilities, and that flexibility is valuable to the extent that the three dimensions carry independent information requiring joint adjustment. In 1X2 forecasting they largely do not. Draw probabilities occupy a narrow band near the base rate and the substantive variation is concentrated in the home-away contrast, so a calibration map has little joint structure to exploit. The fitted vector scaling parameters make this visible: the estimated scales depart from unity by at most 0.041 and the intercepts from zero by at most 0.050, so the fitted map is close to the identity in every dimension.

What this establishes is narrow: at a calibration-set size of roughly four thousand matches, on this problem, the structural distinction between binary-plus-renormalisation and native multiclass calibration is worth 0.0002 in logarithmic score and cannot be resolved from noise on a 4,115-match test period. A larger evaluation period, or a domain whose probability geometry is less constrained, might resolve it, and the direction observed here suggests where such a study would find it.

5.5 The draw problem is a sharpness failure

The literature reviewed in Chapter 2 is close to unanimous that the draw is the outcome football forecasting models handle worst. Section 4.6.3 finds the opposite on the measure the literature usually applies: draw probabilities are the best calibrated of the three outcomes for every real forecaster examined, and the dispersion statistics in the same section explain why.

The two findings are not in conflict, and reconciling them identifies what the received view has been describing. Low calibration error on the draw class is a consequence of a predictive distribution one third as wide as those of the win outcomes. It records that the models are uninformative about draws, not that they are accurate about them.

The difficulty of the draw class is therefore a deficiency of sharpness, not of calibration. The distinction is the one Gneiting and Raftery (2007) identify as fundamental: calibration is the agreement of predicted probabilities with observed frequencies, sharpness is the concentration of the predictive distribution, and the objective is maximal sharpness subject to

calibration. A forecaster that always predicts the base rate is perfectly calibrated and useless. On the draw class, football models approach this degenerate case.

This has a consequence that extends beyond the present study. Calibration diagnostics are structurally incapable of detecting the draw problem, because the problem is not a calibration problem. A study assessing draw forecasting by reliability diagrams or class-wise ECE will conclude that draws are handled well, and will be measuring the wrong thing. Diagnosing the draw problem requires measures sensitive to sharpness - the dispersion of the predicted probability distribution, or a decomposition of a proper scoring rule into calibration and resolution components - rather than calibration measures alone.

5.6 Expected calibration error is unsuitable for model selection

Expected calibration error is the most widely reported scalar calibration measure, and in this study it ranks methods differently from every proper scoring rule. For XGBoost the uncalibrated forecast attains the best ECE and the worst logarithmic score; Dirichlet calibration inverts the ordering.

ECE as conventionally computed summarises the calibration of the predicted-class confidence alone, discarding the remainder of the probability vector, whereas the proper scoring rules evaluate the vector in full. More fundamentally, ECE is not a proper scoring rule, and a forecaster can attain a low value by being reliably unconfident. The calibrated base-rate forecast in this study demonstrates the point, as it attains an ECE superior to that of several substantially more informative forecasts while remaining the worst forecaster in the study by every proper scoring rule.

The same mechanism explains an initially puzzling feature of the market comparison. The de-vigged market forecast outperforms every model on every proper scoring rule while exhibiting a higher ECE than most of them. This is not evidence that the market is poorly calibrated. The market is sharper than the models - its superior accuracy reflects genuine discrimination - and a sharper forecaster has more opportunity to accumulate measured calibration error even when it is the better forecast by every criterion that matters for decision-making.

The concerns raised in the literature about ECE have tended to focus on its sensitivity to binning (Posocco & Bonnefoy, 2021), which is a matter of estimator variance. The problem identified here is different and more serious: it is a matter of what the measure targets. A binning-robust estimator of ECE would still reward unconfidence, and so ECE should not be used to choose between models, as it rewards the unconfident forecaster.

5.7 The market benchmark

The de-vigged market forecast outperforms every model-calibration combination examined on every proper scoring rule, by a margin roughly six times the gain that the best calibration method secured over its uncalibrated counterpart. Calibration narrows the gap without approaching it.

This is the expected result and is consistent with the literature reviewed in Chapter 2. Bookmaker prices incorporate information absent from any feature set constructible from historical results: team news, confirmed lineups, injuries, transfer activity, and the aggregated opinion of a liquid market, all priced by operators with a direct financial incentive to be accurate. A forecast built from publicly available match results and ratings has no mechanism by which to recover this information.

The finding nonetheless carries weight for this study's argument. It establishes that well-calibrated probabilistic forecasting of football outcomes is attainable - the market demonstrates it - and locates the models' deficiency in information rather than in probabilistic technique. Post-hoc calibration addresses how faithfully a model reports what it knows and cannot address what the model does not know. The residual gap to the market is a gap in the covariates, and no calibration method can close it.

5.8 Limitations

A single calibration window. All calibrators were fitted once, on the 2021-2023 seasons. The design is temporally valid but provides no evidence that the observed ranking is stable across time. A walk-forward scheme, advancing the calibration window alongside the test window, would test this directly and would additionally establish whether the separation between parametric and non-parametric maps reported in Section 5.2 holds across differing regimes.

Regularisation structure. The full Dirichlet implementation shrinks W toward the identity under an L_2 penalty rather than applying the off-diagonal scheme of Kull et al. (2019). The sweep reported in Chapter 4 shows the result to be invariant to penalty strength across five orders of magnitude, and vector scaling shows that the limiting case of off-diagonal restriction improves the score slightly. An intermediate constraint would therefore be expected to fall between the two, but this has not been tested directly.

Competition and period scope. The sample pools six European leagues into a single training and test set, over two test seasons. Results are not reported separately by competition, so the consistency across three base learners observed here says nothing about whether the same ranking of calibration methods would hold within any individual league; it is equally possible that the pooled result masks league-specific differences, for instance in draw rates

or competitive balance, that a league-by-league analysis would reveal. Generalisation to other competitions, or to the six competitions studied here taken individually, is not established.

5.9 Directions for further work

The most immediate is the walk-forward calibration design described above, which would convert the study's principal limitation into a result and would test the temporal stability that the present design cannot address.

The second is to characterise the complexity-sample-size frontier more precisely. The learning curve reported here uses one base learner and one set of window sizes. Establishing where the crossover lies for maps of given complexity, and whether it is stable across competitions and periods, would turn a qualitative recommendation into a usable rule.

The third is the decision-theoretic evaluation that this study does not undertake. Because a betting decision is a discontinuous function of a probability - a wager is placed when $p_i o_i$ exceeds a threshold and not otherwise - calibration adjustments small in probability terms can change decisions. Realised return is a noisy way to measure this: with a finite number of bets, the win-or-lose outcome of each individual wager dominates any small, systematic difference between calibration methods, so two methods that make genuinely different decisions can still show similar, or even reversed, returns by chance. Measuring how frequently calibration methods disagree on the decision itself, rather than on the realised return of acting on it, would characterise the economic relevance of calibration without this dependence on the luck of a finite sample of wagers.

5.10 Conclusion

This dissertation set out to determine whether post-hoc probability calibration improves machine-learning forecasts of football outcomes and whether the choice of method matters. It finds that calibration helps reliably but modestly, and that the choice of method matters considerably more than the decision to calibrate - though almost all of that variation lies between the parametric maps and the non-parametric one, not among the parametric maps themselves.

The principal contribution is the identification of calibration-map complexity, relative to calibration-set size, as the variable governing which method should be preferred. It unifies findings that would otherwise appear unrelated: the failure of non-parametric calibration, the ordering of the parametric maps by parameter count at small calibration windows, and the collapse of that ordering into irrelevance at the full window. It rests on a controlled contrast in which complexity varies and nothing else does, and that contrast is significant on one of

the two tree ensembles at a magnitude of 0.0008 - which is to say the mechanism is detectable and the effect is not worth acting on. Because the calibration window in temporally valid football forecasting is bounded by the tension between sample size and representativeness, the constraint is intrinsic to the problem rather than incidental to this design.

Expected calibration error ranks forecasters in systematic contradiction to every proper scoring rule, because it rewards unconfidence, and should not be used for model selection. And the widely reported difficulty of the draw class is a failure of sharpness rather than of calibration, which means the diagnostics conventionally applied to the problem are incapable of detecting it.

None of the calibrated models approaches the de-vigged market forecast, which outperforms them all on every proper scoring rule. That outcome was anticipated and is not a failure of the framework. It places the models' deficiency in what they know rather than in how honestly they report it, and it delimits what post-hoc calibration can be expected to achieve.

References

adamgbor. (2025). *Club Football Match Data (2000-2025).* https://www.kaggle.com/datasets/adamgbor/club-football-match-data-2000-2025. (Kaggle dataset, accessed: 2026-01-31)

Albelali, S., & Ahmed, M. (2025). *Hidden leaks in time series forecasting: How data leakage affects lstm evaluation across configurations and validation strategies.* (arXiv:2512.06932) doi: 10.48550/arXiv.2512.06932

Angelini, G., & De Angelis, L. (2017). Parx model for football match predictions. *Journal of Forecasting*, *36*(7), 795-807.

Angelini, G., & De Angelis, L. (2019). Efficiency of online football betting markets. *International Journal of Forecasting*, *35*(2), 712-721.

Arian, H., Norouzi Mobarekeh, D., & Seco, L. (2024). *Backtest overfitting in the machine learning era: A comparison of out-of-sample testing methods in a synthetic controlled environment.* (SSRN preprint. Published version: Knowledge-Based Systems 304:112477, 2024, doi:10.1016/j.knosys.2024.112477) doi: 10.2139/ssrn.4778909

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., & Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, *26*(4), 641-647. doi: 10.1214/aoms/1177728423

Azhari, H. R., Widyaningsih, Y., & Lestari, D. (2018). Predicting final result of football match using poisson regression model. *Journal of Physics: Conference Series*, *1108*(1), 012066. doi: 10.1088/1742-6596/1108/1/012066

Boshnakov, G., Kharrat, T., & McHale, I. G. (2017). A bivariate weibull count model for forecasting association football scores. *International Journal of Forecasting*, *33*, 458-466.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1-3.

Bunker, R., & Susnjak, T. (2022). The application of machine learning techniques for predicting match results in team sport: A review. *Journal of Artificial Intelligence Research*, *73*, 1285-1322. doi: 10.1613/jair.1.13509

Chen, Z. A. (2025). A hybrid machine learning framework for soccer match outcome prediction: Incorporating bivariate poisson distribution. In *Itm web of conferences* (Vol. 70,

p. 03020). doi: 10.1051/itmconf/20257003020

Cho, H., Yoon, Y., & Lee, S. (2018). Using social network analysis and gradient boosting to develop a soccer win-lose prediction model. *Engineering Applications of Artificial Intelligence*, *72*, 228-240. doi: 10.1016/j.engappai.2018.04.010

Clarke, S., Kovalchik, S., & Ingram, M. (2017). Adjusting bookmaker's odds to allow for overround. *American Journal of Sports Science*, *5*(6), 45-49.

Clarke, S. R., Dixon, M. J., Pope, P. F., & Pleasance, P. (2017). Bookmakers' overround and the favourite-longshot bias: A position paper. *Journal of Sports Analytics*, *3*(4), 255-263. doi: 10.3233/JSA-170222

Constantinou, A. C., & Fenton, N. E. (2012). Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. *Journal of Quantitative Analysis in Sports*, *8*(1). doi: 10.1515/1559-0410.1418

Constantinou, A. C., & Fenton, N. E. (2013). Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal of Quantitative Analysis in Sports*, *9*(1), 37-50.

DeGroot, M. H., & Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *32*(1-2), 12-22. doi: 10.2307/2987588

Diniz, M. A., Izbicki, R., Lopes, D., & Salasar, L. E. (2019). Comparing probabilistic predictive models applied to football. *Journal of the Operational Research Society*, *70*(5), 770-782. doi: 10.1080/01605682.2018.1457485

Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *46*(2), 265-280.

Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Publishing.

Falessi, D., Huang, J., Narayana, L., Thai, J. F., & Turhan, B. (2018). *On the need of preserving order of data when validating within-project defect classifiers.* (arXiv:1809.01510. Published version: Empirical Software Engineering 25(6):4805-4830, 2020, doi:10.1007/s10664-020-09868-x) doi: 10.48550/arXiv.1809.01510

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, *27*(8), 861-874. Retrieved from https://www.sciencedirect.com/science/article/pii/S016786550500303X (ROC Analysis in Pattern Recognition) doi: https://doi.org/10.1016/j.patrec.2005.10.010

Galekwa, R., Radebe, S., Ma, S., Li, Z., & Hao, H. (2024). *A systematic review of machine learning in sports betting: Techniques, challenges, and future directions.* (arXiv:2410.21484) doi: 10.48550/arXiv.2410.21484

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and esti-

mation. *Journal of the American Statistical Association*, *102*(477), 359-378. doi: 10.1198/016214506000001437

Gruber, S. G., & Buettner, F. (2022). Better uncertainty calibration via proper scores for classification and beyond. In *Advances in neural information processing systems 35 (neurips 2022)* (p. 8618-8632).

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th international conference on machine learning (icml 2017)* (Vol. 70, p. 1321-1330).

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. doi: 10.1109/TKDE.2008.239

Hegarty, T., & Whelan, K. (2025). Forecasting soccer matches with betting odds: A tale of two markets. *International Journal of Forecasting*, *41*, 803-820.

Hekler, A., Brinker, T. J., & Buettner, F. (2023). Test time augmentation meets post-hoc calibration: Uncertainty quantification under real-world conditions. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(12), 14856-14864. doi: 10.1609/aaai.v37i12.26735

Hespeler, S. C., Moriano, P., Li, M., & Hollifield, S. C. (2025). *Temporal cross-validation impacts multivariate time series subsequence anomaly detection evaluation.* (arXiv:2506.12183) doi: 10.48550/arXiv.2506.12183

Heuer, A., Müller, C., & Rubner, O. (2010). *Soccer: Is scoring goals a predictable poissonian process?* arXiv preprint arXiv:1002.0797. doi: 10.48550/arXiv.1002.0797

Huang, L., Zhao, J., Zhu, B., Chen, H., & Vanden Broucke, S. (2020). An experimental investigation of calibration techniques for imbalanced data. *IEEE Access*, *8*, 127343-127352.

Hubáček, O., Šourek, G., & Železný, F. (2019). Exploiting sports-betting market using machine learning. *International Journal of Forecasting*, *35*(2), 783-796. doi: 10.1016/j.ijforecast.2019.01.001

Hvattum, L. M., & Arntzen, H. (2010). Using elo ratings for match result prediction in association football. *International Journal of Forecasting*, *26*(3), 460-470.

Jordan, A., Krüger, F., & Lerch, S. (2019). Evaluating probabilistic forecasts with scoringrules. *Journal of Statistical Software*.

Jung, D. H., & Jung, J. J. (2025). Data-driven understanding on soccer team tactics and ranking trends: Elo rating-based trends on european soccer leagues. *PLOS ONE*, *20*(2), e0318485. doi: 10.1371/journal.pone.0318485

Karimov, A., Koshkin, A., Kaplun, D., & Butusov, D. (2025). Domain-driven identification of football probabilities. *Mathematics*, *13*(24), 3976. doi: 10.3390/math13243976

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate poisson models.

Journal of the Royal Statistical Society: Series D (The Statistician), *52*, 381-393.

Kim, P., Lee, S. H., & Seo, J. H. (2024). Evaluating the performance of predictive models using the elo rating system and machine learning algorithms: Empirical evidence of the korean professional football league. *Journal of Digital Contents Society*, *25*(3), 719-732. doi: 10.9728/dcs.2024.25.3.719

Koopman, S. J., & Lit, R. (2015). A dynamic bivariate poisson model for analysing and forecasting match results in the english premier league. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(1), 167-186.

Kull, M., Perelló-Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., & Flach, P. (2019). Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. In *Advances in neural information processing systems 32 (neurips 2019)*. (arXiv:1910.12656) doi: 10.48550/arXiv.1910.12656

Kull, M., Silva Filho, T. M., & Flach, P. (2017a). Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of the 20th international conference on artificial intelligence and statistics (aistats 2017)* (Vol. 54, p. 623-631).

Kull, M., Silva Filho, T. M., & Flach, P. (2017b). Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics*, *11*(2), 5052-5080. doi: 10.1214/17-EJS1338SI

Kumar, A., Liang, P., & Ma, T. (2019). Verified uncertainty calibration. In *Advances in neural information processing systems 32 (neurips 2019)* (p. 3792-3803).

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, *17*(3), 1217-1241.

Kuypers, T. (2000). Information and efficiency: An empirical study of a fixed odds betting market. *Applied Economics*, *32*(11), 1353-1363. doi: 10.1080/00036840050151449

Lee, W.-J., Jhang, H.-J., & Choi, S. H. (2021). Fuzzy study on the winning rate of football game betting. *Advances in Technology Innovation*, *6*(3), 169-178. doi: 10.46604/aiti.2021.6517

Lin, H.-T., Lin, C.-J., & Weng, R. C. (2007). A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, *68*(3), 267-276. doi: 10.1007/s10994-007-5018-6

Liu, H., Gómez, M.-□., Lago-Peñas, C., & Sampaio, J. (2015). Match statistics related to winning in the group stage of 2014 brazil fifa world cup. *Journal of Sports Sciences*, *33*(12), 1205-1213.

Liu, H., Hopkins, W. G., & Gómez, M.-□. (2015). Modelling relationships between match events and match outcome in elite football. *European Journal of Sport Science*. doi: 10 .1080/17461391.2015.1042527

Martino, A., De Santis, E., Baldini, L., & Rizzi, A. (2019). Calibration techniques for binary classification problems: A comparative analysis. In *Proceedings of the 11th international joint*

conference on computational intelligence (ijcci 2019) - volume 1: Ncta (p. 487-495). SciTePress. doi: 10.5220/0008165504870495

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, *12*(4), 595-600.

Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the twenty-ninth aaai conference on artificial intelligence (aaai 2015)* (p. 2901-2907). doi: 10.1609/aaai.v29i1.9602

Newall, P. W. S., & Cortis, D. (2021). Are sports bettors biased toward longshots, favorites, or both? a literature review. *Risks*, *9*, 22.

Niculescu-Mizil, A., & Caruana, R. (2005a). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on machine learning (icml '05)* (p. 625-632). doi: 10.1145/1102351.1102430

Niculescu-Mizil, A., & Caruana, R. (2005b). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on machine learning (icml '05)* (p. 625-632). ACM. doi: 10.1145/1102351.1102430

Nikolić, M., Nikolić, D., Stefanović, M., Koprivica, S., & Stefanović, D. (2025). Mitigating algorithmic bias through probability calibration: A case study on lead generation data. *Mathematics*, *13*(13), 2183. doi: 10.3390/math13132183

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., & Tran, D. (2019). Measuring calibration in deep learning. In *Ieee/cvf conference on computer vision and pattern recognition workshops (cvprw)* (p. 38-41).

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., … Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in neural information processing systems 32 (neurips 2019)*.

Pachur, T., & Biele, G. (2007). Forecasting from ignorance: The use and usefulness of recognition in lay predictions of sports events. *Acta Psychologica*, *125*, 99-116.

Peeters, T. (2018). Testing the wisdom of crowds in the field: Transfermarkt valuations and international soccer results. *International Journal of Forecasting*.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), *Advances in large margin classifiers* (p. 61-74). Cambridge, MA: MIT Press.

Posocco, N., & Bonnefoy, A. (2021). Estimating expected calibration errors. In *Artificial neural networks and machine learning - icann 2021* (Vol. 12894, p. 139-150). Springer. doi: 10.1007/978-3-030-86380-7_12

Rodrigues, F., & Pinto, M. F. (2022). Prediction of football match results with machine learning. In *Procedia computer science* (Vol. 204, p. 463-470). doi: 10.1016/j.procs.2022.08.057

Saribekyan, G., & Yarovoy, N. (2025). Football prediction model based on the teams' elo ratings and scoring indicators. *Journal of Journalism and Media Management*, *1*(1), 1-4. doi: 10.61440/JJMM.2025.v1.02

Sarmento, H., Marcelino, R., Anguera, M. T., Campaniço, J., Matos, N., & Leitão, J. C. (2014). Match analysis in football: A systematic review. *Journal of Sports Sciences*. doi: 10.1080/02640414.2014.898852

Spann, M., & Skiera, B. (2009). Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting*, *28*(1), 55-72. Retrieved from https://EconPapers.repec.org/RePEc:jof:jforec:v:28:y:2009:i:1:p:55-72

Stekler, H. O., Sendor, D., & Verlander, R. (2010). Issues in sports forecasting. *International Journal of Forecasting*, *26*, 606-621.

Stübinger, J., Mangold, B., & Knoll, J. (2020). Machine learning in football betting: Prediction of match results based on player characteristics. *Applied Sciences*, *10*(1), 46. doi: 10.3390/app10010046

Transfermarkt GmbH & Co. KG. (2025). *Transfermarkt: Football transfers, market values and statistics.* Retrieved from https://www.transfermarkt.com (Accessed: 2025-02-01)

Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., & Schön, T. B. (2019). Evaluating model calibration in classification. In *Proceedings of the 22nd international conference on artificial intelligence and statistics (aistats 2019)* (Vol. 89, p. 3459-3467).

Walsh, C., & Joshi, A. (2024). *Machine learning for sports betting: should model selection be based on accuracy or calibration?* (arXiv:2303.06021) doi: 10.48550/arXiv.2303.06021

Wheatcroft, E. (2021). Evaluating probabilistic forecasts of football matches: the case against the ranked probability score. *Journal of Quantitative Analysis in Sports*, *17*(4), 273-287. doi: 10.1515/jqas-2019-0089

Wilkens, S. (2026). Can simple models predict football - and beat the odds? Lessons from the German Bundesliga. *Journal of Sports Analytics*, *12*. doi: 10.1177/22150218261416681

World Football Elo Ratings. (2026). *World Football Elo Ratings.* https://www.eloratings.net/. (Accessed: 2 Feb 2026)

Zadrozny, B., & Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the eighteenth international conference on machine learning (icml '01)* (p. 609-616). Morgan Kaufmann.

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth acm sigkdd international conference on knowledge discovery and data mining (kdd '02)* (p. 694-699). ACM. doi: 10.1145/775047.775151

Zhao, Z., Liu, H., & Prakash, B. A. (2025). *Tackling time-series forecasting generalization via mitigating concept drift.* (arXiv:2510.14814) doi: 10.48550/arXiv.2510.14814

Štrumbelj, E. (2014). On determining probability forecasts from betting odds. *International*

Journal of Forecasting, *30*, 934-943.