STAT0035
Using Machine Learning Methods to Identify Value in Football Betting Markets

Word Count: 9667

Candidate LYFP4

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Science
of
University College London.

Department of Statistical Science
University College London

April 27, 2025

I, Candidate LYFP4, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Predicting football match outcomes is a complex task influenced by numerous factors, including team performance, historical trends, and contextual variables. This study presents a comprehensive modelling framework for predicting match results across three outcomes: home wins, away wins, and draws. The framework will be used to inform a betting strategy.

The study employs a dual-model approach, integrating both logistic regression and machine learning models, to address the challenges of multi-class classification, particularly the underrepresented draw category. Historical data from the English Premier League (2015–2023) serves as the foundation for training and testing the models. Model performance is evaluated from two main perspectives. Firstly, the models are evaluated with statistical measures including accuracy, Brier scores, and log scores, and compared against bookmaker odds as a benchmark. Next, the model is evaluated with economical measures, through the implementation of a betting strategy. The betting strategy uses both edge-based and Sharpe Ratio-based approaches. The profitability of these strategies is analysed, with a focus on scenarios where the model's predicted probabilities diverge from bookmaker-implied probabilities.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Jim Griffin, for his guidance and support throughout this project. I am grateful for our fruitful weekly discussions and engaging conversations about football, which were instrumental in shaping my ideas.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Preface

Football, often referred to as "the beautiful game," has captured the hearts of millions worldwide, with 1.5 billion viewers tuning into the 2022 FIFA World Cup Final (FIFA, 2023). Association football also garners high engagement levels - The English Premier League has become one of the UK's most successful global exports, broadcasting to 189 out of 193 UN member states and reaching 900 million homes worldwide (Premier League, 2024). This extensive reach has translated into an impressive following, with 1.87 billion people interacting with the Premier League at least weekly through various media channels.

1.2 Statistics in Football

The use of statistical science in football has become increasingly prevalent in the past decade, becoming an increasing part of both casual fan engagement and professional team management.

On a casual level, fantasy football has emerged as a popular platform for fans to engage in statistical analysis. Fantasy football managers work with a restricted pool of in-game money, used to select a team of football players who are subsequently awarded points based on their real-life performances. Managers usually create a mini-league, to compete amongst friends and family. Fantasy football managers attempt to incorporate a wide range of factors,

including fixture difficulty ratings, predicted playing time, and Expected Goals (xG) in their team selections. The unpredictability of individual football performances poses an exciting challenge for many fantasy managers.

At the professional level, football clubs are increasingly investing in predictive data analysis to gain a competitive edge.

For example, Liverpool FC has demonstrated the potential of data-driven recruitment. Their approach, spearheaded by data scientist Dr. Ian Graham, utilises advanced statistical models to identify undervalued talent. One of the most successful outcomes of this strategy was the signing of Mohamed Salah in 2017. The decision was based on a "Possession Value" model, which evaluated Salah's comprehensive game contribution beyond just goal-scoring. Despite initial skepticism, the club chose to proceed with the signing of Salah, which proved to be a shrewd decision (Bate, 2024). Salah has since become Liverpool's fourth all-time top goalscorer and was instrumental in ending the club's 30-year wait for an English league title.

Other clubs like Brentford FC have also achieved consistent success through data-driven methods. Brentford FC's "Moneyball" approach has helped them uncover undervalued players. For example, they signed players such as striker Ollie Watkins and goalkeeper David Raya, both of whom have gone on to be prolific players at other clubs. Brentford's rise from League One to the Premier League demonstrates how the effective use of statistics can allow smaller clubs to compete against wealthier rivals (Lewis and Nabbi, 2023).

As more clubs adopt these methods, understanding and effectively implementing statistical science is becoming crucial for success in modern football management. This also underpins the high potential of statistical science in football.

1.3 Football Betting

Betting in the United Kingdom has a long history, dating back to the early 20th century. Initially, it was characterised by informal wagers among friends and underground bookmakers (The Footy Tipster, 2023). The landscape of betting underwent a significant transformation with the introduction of the Betting and Gaming Act of 1960.

This legislation not only legalised betting shops but also ignited an unprecedented boom in the commercial gambling sector. However, the rapid expansion came with significant drawbacks, attracting organised criminal elements. The Gaming Act of 1968 sought to remove the criminality associated with commercial gambling by restricting gambling activity to licensed premises, setting the premise for modern sports betting (House of Commons, 2012).

Today, sports betting is a popular activity in the United Kingdom, with 17% of adults placing at least one sports bet in 2019. (Department for Digital, Culture, Media & Sport, 2020). In 2024, the sports betting industry was estimated to be worth more than £3.3 billion (Statista Research Department, 2024). Bookmakers, who are central to shaping betting markets, employ sophisticated methods to set odds that reflect their predictions of match outcomes. These odds are derived from comprehensive databases, real-time information, and various factors including team/player performance, injuries, historical data, and even public sentiment.

1.4 Aims

This project aims to develop a predictive model for the outcomes of football matches, which will inform a profitable betting strategy. Bookmaker odds serve as an industry benchmark against which to evaluate the model's performance. Therefore, the profitability of the betting strategy will serve as the main measure of the performance of the predictive model.

Chapter 2

Literature Review

2.1 Poisson Distributions in Football

Several projects have attempted to model football match results, with one of the most notable approaches being the use of independent Poisson distributions to model the number of goals scored. Maher (1982) pioneered this method, introducing a framework that incorporates parameters for teams' attacking and defensive strengths, as well as home advantage. His model assumes that the number of goals scored by each team follows independent Poisson distributions, providing a probabilistic foundation for predicting match outcomes.

Maher's analysis of 12 separate leagues demonstrated that separate attack and defense parameters for each team were necessary to capture team-specific dynamics. However, home and away differences could be modelled using a single factor, simplifying the structure of the model. This approach laid the groundwork for subsequent advancements in football match modelling.

Dixon and Coles (1997) expanded on Maher's work by addressing two key limitations of the original model. First, they introduced a dependence parameter to account for the departure from independence in low-scoring games. Second, they incorporated a parameter to capture seasonal variations in team form, acknowledging that performance levels fluctuate over the course of a season. These modifications resulted in slight but meaningful improvements to the model's predictive performance.

Despite these advancements, Poisson-based models have inherent limitations. They rely heavily on the assumption that goal-scoring events are independent and identically distributed, which may not hold true in practice. Additionally, these models often struggle to account for the rarity of draw outcomes, which are underrepresented in football data. These limitations have prompted researchers to explore alternative modelling approaches that offer greater flexibility and predictive power.

2.2 Result-Based Modelling

More recent literature has generally favoured the use of result-based modelling, especially for the purposes of creating a profitable betting model. Using a result-based model also allows the integration of more advanced metrics and machine learning methods which will be further explored.

Goddard (2005) integrates both goal-based and result-based models in their research. He posits that while goals-based models may draw on more extensive data, results-based models benefit from clearer specifications and may be equally effective for forecasting match outcomes. Tsokos et al. (2019) builds on the result-based approach by specifying a Bradley-Terry model. They investigated several extensions of the original Bradley-Terry formulation, allowing team strengths to vary over time and depend on various features. Tsokos et al. (2019) estimate the parameters of these Bradley-Terry extensions by maximising the log-likelihood or a penalised version of it.

While Tsokos et al. (2019) focus on more traditional statistical models, they acknowledge the growing trend towards machine learning approaches in sports analytics. Recent literature has seen an increasing popularity of machine learning techniques for sports prediction.

2.3 Machine Learning in Football Prediction

Traditional statistical models, such as logistic regression or the Bradley-Terry model, often rely on linear assumptions and predefined features. The assumption of linearity is not necessarily one that should be made. Modern projects

have explored the utilisation of machine learning techniques, such as cross-validation, in traditional statistical models. These methods have been shown to improve predictive accuracy. In addition, machine learning models like Random Forests, XGBoost, and neural networks have been shown to be better at capturing non-linear relationships, and possibly more useful in the realm of modelling football results.

A study by Hubáček et al. (2019) highlights the potential of machine learning methods in sport prediction. They employ a convolutional neural network that leverages player-level statistics, using the convolution layer to aggregate individual player features into team-level predictions. The study also underscored the importance of feature engineering and domain-specific knowledge in optimising machine learning workflows.

Baboota and Kaur (2019) also utilises machine learning methods and applies them directly to football result prediction. They found that gradient boosting produced the best predictive results.

Similarly, Fahey-Gilmour et al. (2019) applies machine learning techniques to Australian football, discovering that a generalised linear model with elastic net regularisation (glmnet) performed best for predicting match outcomes. Their model achieved 73.3% accuracy on the test set, rivaling bookmaker predictions. This finding highlights the potential of regularised linear models in sports prediction tasks, particularly when dealing with high-dimensional feature spaces. Both approaches offer advantages over traditional statistical methods, particularly in their ability to handle large numbers of features and capture intricate patterns in the data. Therefore, this project will employ machine learning methods to conduct statistical analysis.

2.4 Betting

The ultimate aim of this project is to develop a profitable sports betting system using machine learning techniques.

Dixon and Coles (1997) developed a statistical model for predicting football match outcomes with the aim of identifying profitable betting opportunities. Their approach focused on exploiting potential inefficiencies in the betting market. The authors tested their model's performance in a simulated betting scenario, developing a strategy where bets were placed only when the model's predicted probability for a particular outcome exceeded the bookmaker's implied probability by a certain threshold. They found that betting on away wins was generally unprofitable. Interestingly, their simple strategy found some success. However, they also state that the addition of more covariates could result in a more refined model. Given the vast increase in available football statistics and metrics since 1997, there is much more information available to incorporate in the models today.

Rue and Salvesen (2000) developed a sophisticated Bayesian model for predicting football match outcomes and devised a corresponding betting strategy. Their approach utilises dynamic modelling techniques to estimate team strengths, allowing these parameters to evolve over time. The model incorporates various factors influencing match results, including home advantage, attack and defense strengths of teams, and recent form. For their betting strategy, Rue and Salvesen propose a method that compares their model's predicted probabilities with bookmakers' odds to identify potentially profitable bets. They introduce a utility function that balances the expected return of a bet against its associated risk, allowing bettors to adjust their strategy based on individual risk preferences. The authors suggest focusing on bets with the highest expected utility rather than simply betting on all favorable odds. Their simulations indicated that this strategy could potentially yield positive returns, particularly when applied to less popular betting markets where bookmakers' odds might be less efficient.

Building on this foundation, subsequent research has explored more sophisticated betting strategies. As the sports betting market has evolved, so too have the approaches to identifying profitable opportunities. One key area

of development has been the recognition that simply improving predictive accuracy may not be sufficient to overcome bookmakers' edge. This has led researchers to explore novel methods for exploiting market inefficiencies and developing more nuanced betting strategies.

Hubáček et al. (2019) used a similar portfolio optimisation strategy for National Basketball Association (NBA) games and achieved positive profits. Their strategy also aimed to balance expected returns and variance using principles from modern portfolio theory. Bets are distributed to maximise the Sharpe Ratio, which measures profitability relative to risk, ensuring that the bettor allocates their budget across outcomes with the highest potential market inefficiency. Another interesting finding from their study was that a high correlation between the bettor's predicted probabilities and bookmaker odds would be detrimental to the bettor. Every bookmaker has an edge—their profit margin—and even if the bettor's model was highly accurate, if it perfectly coincided with the bookmaker's model then they would undoubtedly lose money. To counter this, the authors introduced a decorrelation parameter, which modifies the model's loss function to penalise alignment with the bookmaker's implied probabilities, encouraging predictions that are accurate but independent of the bookmaker's odds.

Chapter 3

Sources and Data

3.1 Sources

The data used in this project is sourced from multiple platforms that provide extensive football-related datasets. The primary sources are **FBref**, which was accessed via the R package **worldfootballR**, and **football-data.co.uk**, which offers detailed odds data from various bookmakers. The project is also supplemented with Elo ratings from **clubelo.com** to enhance the predictive modeling process.

3.2 Data

3.2.1 Match Results and Team Performance Metrics

The match results and team performance metrics were obtained using the **worldfootballR** package. This package allows access to match data from most major football data providers, covering detailed match-level data from leagues worldwide. For this study, the English Premier League match data between 2015 and 2023 was extracted.

Key features included:

- **Match outcomes:** Results were categorised into Home Win, Away Win, or Draw.
- **Goals scored:** Separate counts for home and away teams.

- **Expected Goals (xG):** Provided only after the 2017/18 season
- **Match dates and venues:** To allow for time-series analysis and venue-specific adjustments.

3.2.2 Bookmaker Odds

Odds data from **football-data.co.uk** encompassed the probabilities implied by major bookmakers. The data provided also included odds for different bets such as the 2.5 goals bet and the Asian handicap. However, for the purposes of this study, only the odds for the pure result will be used. Conveniently, the dataset provides the maximum odds available across all the bookmakers starting from season 2019/20. Hence, these maximum odds will be used for the purposes of testing.

These odds were converted into implied probabilities for Home Win, Away Win, and Draw. As expected, the sum of these implied probabilities was always more than 1. This was the bookmaker's margin. To adjust for the bookmaker's margin, the implied probabilities are scaled so their sum equals 1. This is achieved by dividing each implied probability by the sum of all implied probabilities. The adjusted probabilities can be expressed as:

$$P_{\text{adjusted}}(Outcome) = \frac{P_{\text{implied}}(Outcome)}{\sum P_{\text{implied}}(Outcome)},$$

3.2.3 Elo Ratings

Elo ratings, originally developed for chess players by Professor Arpad Elo (1978), are a dynamic ranking system that measures a team or player's relative performance based on past results. This system has since been adapted for various sports, including football, due to its effectiveness in capturing team strength over time. The Elo rating formula for football can be expressed as:

$$R_n = R_o + K \times (W - W_e) \tag{3.1}$$

where:

- R_n is the new rating
- R_o is the pre-match rating
- K is a weight constant based on the tournament type
- W is the match result (1 for win, 0.5 for draw, 0 for loss)
- W_e is the expected result based on pre-match ratings

The Elo rating system dynamically adjusts team ratings after each match, with the magnitude of change depending on the relative strength of the opponents and the match outcome. For instance, if a weaker team defeats a much stronger team, the resulting change in ratings will be larger for both teams compared to a scenario where the weaker team defeats an opponent of similar strength. This adaptability makes Elo ratings particularly useful for evaluating team performance over time.

Elo ratings offer several advantages over traditional ranking methods. First, they account for the strength of opponents, providing a more nuanced evaluation than simple win-loss records. Second, the system has strong predictive power, making it a valuable tool for forecasting match outcomes. Third, Elo ratings are objective and based solely on performance, which helps to mitigate potential biases that may arise in subjective ranking systems.

For this study, Elo ratings were obtained from **clubelo.com**, a reputable source for football team rankings. The ratings provided by **clubelo.com** go beyond basic match results by incorporating additional factors such as home advantage and goal difference. This ensures that the ratings offer a more comprehensive and accurate measure of team strength, making them highly suitable for use in this analysis.

3.2.4 Feature Engineering

Additional features were engineered from the raw data to improve model performance:

- **xG Moving Averages:** To capture trends in team performance across recent matches, the 5-game moving average for xG was calculated sepa-

rately for home and away games. This separation is useful because teams often perform differently depending on whether they are playing at home or away, and go through patches of form which the moving averages aim to reflect.

- **Goal Differences:** To reflect relative scoring ability, the goal difference (goals scored minus goals conceded) was also calculated separately for home and away performances. This feature provides insight into a team's offensive and defensive capabilities.

3.2.5 Data Cleaning and Standardisation

To ensure consistency, team names were standardised across different datasets using regular expressions. This step resolved discrepancies in naming conventions, such as abbreviations (`Utd` to `United`) or alternative spellings (`Nott'ham Forest` to `Nottingham Forest`).

Chapter 4

Modelling

This chapter outlines the methodological framework employed in the modelling process, focusing on the development and evaluation of predictive models for football match outcomes. The models are assessed from both statistical and financial perspectives, with this chapter primarily addressing the statistical evaluation.

Match outcomes (Home Win, Draw, or Away Win) were predicted using features such as moving averages of expected goals (xG), pre-match Elo ratings, and contextual factors like match venue. As a general rule of thumb, five seasons of data were used to predict the matches for the testing season. Three seasons of the English Premier League were tested.

The entire modelling pipeline is implemented in `R`, utilising packages such as `xgboost` for model training, `caret` for hyperparameter tuning, and `dplyr` for data manipulation. Reproducibility is ensured through consistent random seeds and a modular code structure.

Several approaches were explored during the development of the predictive models.

4.1 Single-Stage Models

4.1.1 Initial Approach: Bradley-Terry Model

The first attempt considered the Bradley-Terry model, a classic method for pairwise comparisons. The Bradley-Terry model (Bradley and Terry, 1952)

was originally developed to estimate the relative strengths of competing teams or individuals through pairwise comparisons. The model assumes that the probability of one team beating another is proportional to their relative strengths.

The probability that Team i beats Team j is given by

$$P(\text{Team } i \text{ beats Team } j) = \frac{\pi_i}{\pi_i + \pi_j}, \tag{4.1}$$

where:

- π_i and π_j represent the positive exponential strength parameters for Teams i and j, respectively. Higher values indicate stronger teams.

4.1.1.1 Estimating the Strength Parameters

The strength parameters (π_i) are typically estimated using maximum likelihood estimation (MLE). Given a set of observed match outcomes, we first define:

- w_{ij} as the number of matches where Team i beats Team j,
- w_{ji} as the number of matches where Team j beats Team i.

Under the assumption that pairwise outcomes are independent, the likelihood function for the Bradley-Terry model becomes

$$L(\boldsymbol{\pi}) = \prod_{i<j} \left(\frac{\pi_i}{\pi_i + \pi_j} \right)^{w_{ij}} \left(\frac{\pi_j}{\pi_i + \pi_j} \right)^{w_{ji}}, \tag{4.2}$$

where the product is taken over all unordered pairs of teams.

To simplify the estimation process, we take the logarithm of the likelihood function to obtain the log-likelihood:

$$\ell(\boldsymbol{\pi}) = \sum_{i<j} \left[w_{ij} \log \pi_i + w_{ji} \log \pi_j - (w_{ij} + w_{ji}) \log(\pi_i + \pi_j) \right]. \tag{4.3}$$

The strength parameters π_i are then estimated by maximising this log-likelihood function. An identifiability constraint (e.g., $\sum_i \pi_i = 1$ or fixing one

π_i) is imposed to ensure a unique solution.

4.1.1.2 Limitations of the Bradley-Terry Model

While the Bradley-Terry model is effective for estimating relative strengths, it has several limitations, particularly in the context of association football.

Firstly, the model does not account for draws, which are a significant outcome in football. Approximately 23.5% of games between the 2014/15 and 2022/23 Premier League seasons resulted in draws, making this limitation relevant. Previous research has proposed modifications to the Bradley-Terry framework to incorporate the probability of draws (e.g., by modeling outcomes as multinomial ordered outcomes (Tsokos et al., 2019)), but this adds complexity to the model.

Next, the model assumes independence across all comparisons, which may not hold in football. Teams often experience fluctuations in performance, with periods of good or poor form, and additional factors such as time since the last fixture, injuries and the whole context around the fixture can influence outcomes. The standard Bradley-Terry model does not account for these dynamic, contextual factors.

Lastly, for high-dimensional datasets (many teams and matches), estimating the strength parameters can be computationally challenging, and the model may struggle to scale effectively.

Due to these limitations, alternative approaches that incorporate additional factors such as team form, injuries, and other contextual variables may provide a more nuanced understanding of team performance and match outcomes.

4.1.2 Single-Stage `glmnet` Model (1S-GLM)

The model will be referred to as the Single-Stage `glmnet` Model (1S-GLM).

A multinomial logistic regression model with a penalised maximum likelihood has shown promise in predictive modeling for sports, such as Australian football (Fahey-Gilmour et al., 2019). `glmnet` is the package used to fit such

a model.

First, a multinomial logistic regression model is fitted. Following which, the package uses LASSO and ridge regression to form an elastic net penalty, which is applied to the log-likelihood function. The elastic net combines the strengths of both L1 (LASSO) and L2 (ridge regression) penalties, offering a balanced approach to coefficient shrinkage and feature selection. The L1 component, characterised by the absolute value of coefficients, promotes sparsity by shrinking some coefficients precisely to zero, effectively performing variable selection by eliminating less influential predictors from the model. Conversely, the L2 component, which uses the squared coefficients, implements a more conservative shrinkage that retains all variables while effectively addressing multicollinearity by proportionally reducing the impact of correlated features. By integrating these complementary penalties, elastic net manages the limitations of each individual approach. The elastic net handles groups of correlated variables more effectively than a pure LASSO penalty, while maintaining the feature selection capability that the ridge regression penalty lacks. The elastic net's blended penalty is controlled by the hyperparameter $\alpha \in [0, 1]$, which determines the relative contribution of each penalty type (Equation 4.5).

The multinomial logistic regression model extends the binomial framework to handle K outcome classes ($G = \{1, \ldots, K\}$). For observation i, the probability of class k is:

$$p_{ik} = \Pr(G_i = k \mid X = x_i) = \frac{\exp\left(\beta_{0k} + \beta_k^T x_i\right)}{\sum_{\ell=1}^{K} \exp\left(\beta_{0\ell} + \beta_\ell^T x_i\right)}, \tag{4.4}$$

where:

- β_k: Coefficient vector for class k
- x_i: Feature vector for observation i
- β_{0k}: Intercept term for class k

Following which, the elastic net penalised negative log-likelihood is:

$$\ell(\{\beta_{0k}, \beta_k\}_{k=1}^K) = -\left[\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} y_{ik}(\beta_{0k} + x_i^T\beta_k) - \log\left(\sum_{\ell=1}^{K} e^{\beta_{0\ell}+x_i^T\beta_\ell}\right)\right)\right] + \lambda\left[(1-\alpha)\frac{\|\beta\|_F^2}{2} + \alpha\sum_{j=1}^{p}\|\beta_j\|_q\right] \quad (4.5)$$

where:

- y_{ik}: Indicator variable (1 if observation i belongs to class k, 0 otherwise)
- λ: Regularisation strength ($\lambda \geq 0$), controlling overall penalty magnitude. Larger λ increases the weight of the elastic net.
- $\alpha \in [0, 1]$: Elastic net mixing parameter:
  - $\alpha = 1$: Pure L1 penalty (LASSO)
  - $\alpha = 0$: Pure L2 penalty (ridge regression)
  - $0 < \alpha < 1$: Hybrid of L1/L2 penalties — balances variable selection and grouped-correlation handling.
- $\|\beta\|_F^2$: L2 Penalty, Frobenius norm squared ($\sum_{k=1}^K \sum_{j=1}^p \beta_{kj}^2$) — the ridge regression penalty shrinks coefficients toward zero.
- $\|\beta_j\|_q$: L1 Penalty for variable j across all K classes, with $q \in \{1, 2\}$:
  - $q = 1$: $\|\beta_j\|_1 = \sum_{k=1}^K |\beta_{jk}|$ — applies LASSO penalty per coefficient, allowing outcome-specific sparsity.
  - $q = 2$: $\|\beta_j\|_2 = \sqrt{\sum_{k=1}^K \beta_{jk}^2}$ — applies grouped-LASSO penalty. Coefficients for variable j are zero or non-zero simultaneously across all classes.

Cross-validation was used to optimise α and λ, the penalty mixing parameters. The `caret` package facilitated the optimisation process by performing a grid search over a predefined range of α and λ values. Specifically, a 10-fold cross-validation approach was implemented, where the dataset was partitioned

into 10 subsets, and the model was iteratively trained on 9 subsets and validated on the remaining one. The optimal α and λ values were selected based on the lowest average log loss across the validation folds.

The optimal values for the mixing parameters varied over the 3 seasons which were tested. Summarised in the table below are the values of α and λ chosen.

Table 4.1: Optimal Mixing Parameters for 1S-GLM

Season	α	λ
2020/21	0.56	0.010
2021/22	0.89	0.022
2022/23	0.56	0.022

Results above suggest a slight preference for L1 (LASSO) penalties, with α generally putting more weight on the L1 penalty, implying sparse feature selection may be advantageous. However, mid-range α values reflect the benefit of L2 (ridge regression) penalites, as multicollinear team-quality indicators (e.g., Elo ratings, xG) benefit from L2 shrinkage to stabilise coefficients and mitigate overfitting.

It was found that the grouped LASSO penalty (q=2) produced marginally better prediction results than the standard LASSO (q=1). Conceptually, the grouped LASSO penalty may be more suitable for this purpose, as it enforces sparsity at the variable level across all outcome classes simultaneously. This ensures that a predictor is either retained for all classes (Home Win, Draw, Away Win) or excluded entirely, implying that influential variables might holistically impact match outcomes rather than selectively affecting only specific results.

4.1.2.1 Evaluation and Limitations

The model consistently predicted a low probability of draws, not once indicating that a draw was the most likely outcome. To attempt to fix this, a weighting scheme was applied, assigning a higher weight to draws during training. However, that further reduced the quality of predictions. This limitation is a

common feature across various predictive models and even bookmaker odds, which tend to favor decisive outcomes over ties.

This limitation could stem from its regularised regression framework - draw outcome predictions suffered from excessive penalisation, with elastic net regularisation disproportionately nullifying feature coefficients and reducing draw probabilities to intercept-only baselines. This phenomenon arose from the relative rarity of draws and the L1 penalty's aggressive pruning of infrequent predictors. These constraints rendered draw predictions less sensitive to the variables of the model, functioning as static background probabilities.

Table 4.2: Model Coefficients (2020/21 Season)

Feature	**HomeWin** β	**AwayWin** β
Home xG	0.100	-0.110
Away xG	-0.039	0.233
Home Elo	0.0029	-0.0033
Away Elo	-0.0020	0.0035

Standard error estimates for these coefficients are not directly available due to the use of penalised multinomial regression via `glmnet`, which does not produce standard errors in the traditional sense. While bootstrapping could approximate them, it is computationally intensive and not essential here, as the primary aim is model interpretation and relative feature importance rather than statistical inference.

As expected, the model produced differing feature weights for expected goals (xG) metrics. Away team xG demonstrated disproportionately strong impacts, with 2020/21 coefficients showing $\beta_{\text{AwayWin\_AwayxG}} = 0.233$ versus $\beta_{\text{HomeWin\_HomexG}} = 0.100$, suggesting away attacking quality was 2.33 times more predictive of away wins than home xG was of home victories. This asymmetric weighting contradicts conventional football wisdom where home advantage typically amplifies home team metrics, indicating that the strength of the away team may have a bigger part to play in the results of a game than the strength of the home team.

The elastic net implementation exhibited notable stability in Elo rat-

ing impacts, with home team coefficient variations remaining minimal across seasons ($\Delta\beta_{\text{HomeElo}} < 0.0001$ year-on-year). This consistency in parameter estimation, coupled with the stronger relative weights for away team metrics in both xG ($\beta_{\text{AwayxG}} = 0.233$ vs $\beta_{\text{HomexG}} = 0.100$) and Elo coefficients ($\beta_{\text{AwayElo}} = 0.0035$ vs $\beta_{\text{HomeElo}} = 0.0029$), suggests conventional home advantage effects may be less deterministic than widely assumed. Rather than home team strength acting as the dominant predictor, the model implies match outcomes are more contingent on away team quality, which aligns with what was suggested above.

The model's constrained draw predictions and asymmetric xG weights ultimately reflect the conflict between parsimony and complexity in regularised regression. While good for overall predictions and the prevention of overfitting, the 1S-GLM's feature selection proved too aggressive for rarer outcomes, which in this case are draws. Future implementations might benefit from outcome-specific modelling to better capture the different nuances for different outcomes.

4.1.3 Single Stage-XGBoost Model (1S-XGB)

The **Single-Stage XGBoost Model (1S-XGB)** employs XGBoost, a gradient-boosted decision tree algorithm known for its strength in capturing non-linear relationships and interactions in data. This is particularly advantageous in football analytics, where match outcomes depend on complex, interdependent factors. The 1S-XGB treats match outcomes (Home Win, Draw, Away Win) as a multiclass classification problem, estimating probabilities for each class using a softmax-transformed score-based objective. Below, we first demonstrate an example of a decision tree, before showing its mathematical formulation, optimisation process, and implementation process.

4.1.3.1 Example XGBoost Decision Tree Process

This example demonstrates how 1S-XGB incrementally builds an ensemble model through gradient boosting. Each tree corrects residuals from previous

iterations, with splits optimised to minimise the regularised objective function (Equation 4.10). In this example, the maximum depth of the tree is 3.

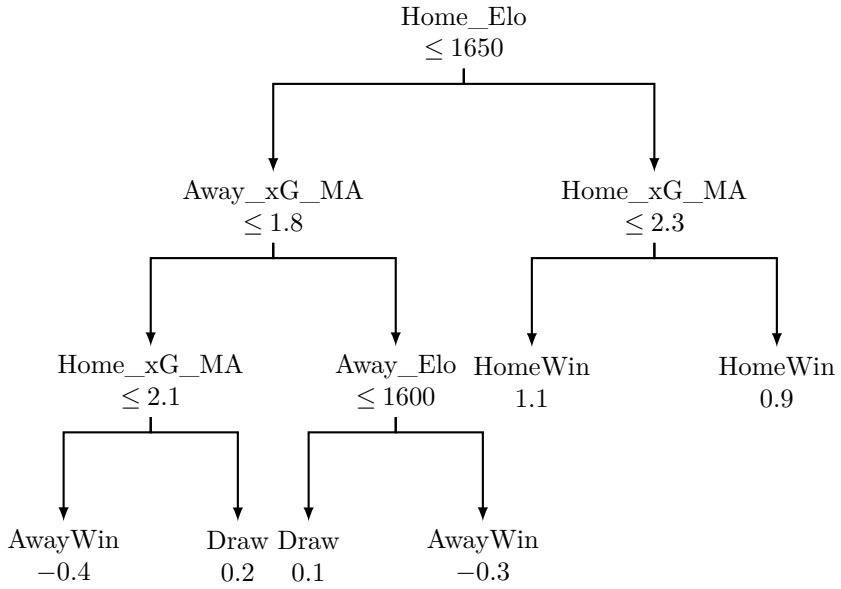

Figure 4.1: Example Tree Structure

The illustrated tree represents a single iteration ($t = 1$) in the XGBoost ensemble process, as defined by Equation 4.13. Over $T = 50$ boosting rounds, the model iteratively constructs such trees, refining predictions by aggregating their outputs.

At each iteration, the model updates the total score for class c by adding the output of the newly built tree. The cumulative score is computed as according to the scoring function (Equation 4.13).

Each new tree is constructed to minimise the regularised objective function (Equation 4.9). To achieve this, XGBoost computes the gradients and Hessians based on the previous iteration's predictions $\hat{y}^{(t-1)}$:

- The gradient: $g_i = \partial_{\hat{y}^{(t-1)}} \ell(y_i, \hat{y}^{(t-1)})$, representing the direction of steepest descent.
- The Hessian: $h_i = \partial^2_{\hat{y}^{(t-1)}} \ell(y_i, \hat{y}^{(t-1)})$, which provides a second-order approximation of curvature.

These values guide how the tree splits the data to reduce the overall loss.

After 50 boosting rounds, the final predicted probabilities are computed using the softmax function (Equation 4.14) where each leaf value (e.g., 1.1 or -0.4) represents an individual tree's contribution before learning rate scaling. This iterative process continues until a stopping criterion, such as a maximum number of trees or early stopping, is met.

4.1.3.2 Modelling Process: Tree Construction & Boosting

After establishing the concept of the tree, we may proceed with the construction of the model. The XGBoost training process builds an ensemble of gradient-boosted decision trees through an iterative and additive approach. Each iteration in the process contributes to refining the predictions and improving model performance.

1. **Initial Prediction:**

   The process starts with a base score, typically set to 0.5 for classification tasks, before any trees are built. This serves as the initial estimate for all instances:

   $$\hat{y}_i^{(0)} = \text{Base Score} \tag{4.6}$$

2. **Formulation of the Objective Function:** Before proceeding with the creation of the tree, the loss and objective functions must be defined. These functions will influence the way the branches of the tree are split. The objective function in XGBoost combines the multiclass logarithmic loss with a regularisation term to both measure classification performance and control model complexity. The model's classification performance is measured using the multiclass logarithmic loss (mlogloss):

   $$\text{mlogloss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\log(p_{ic}), \tag{4.7}$$

   with:

   - N representing the number of instances in the dataset,
   - p_{ic} is the predicted probability of instance i belonging to class c,

- y_{ic} being a binary indicator that equals 1 if instance i belongs to class c, and 0 otherwise.

To mitigate overfitting and control model complexity, a regularisation term is incorporated. For each tree g_{tc}, the regularisation term is defined as:

$$\Omega(g_{tc}) = \gamma T_{tc} + \frac{1}{2}\lambda \sum_{j=1}^{T_{tc}} w_{tcj}^2 + \alpha \sum_{j=1}^{T_{tc}} |w_{tcj}|, \tag{4.8}$$

where:

- T_{tc} is the number of leaves in the t-th tree for class c,
- w_{tcj} is the weight (score) of the j-th leaf in the t-th tree,
- γ is the minimum loss reduction required to perform a split,
- λ is the coefficient for L2 regularisation (ridge penalty),
- α is the coefficient for L1 regularisation (lasso penalty).

The overall objective function that the model seeks to minimise is a sum of the multiclass log loss and the regularisation penalties across all trees:

$$\mathcal{L}(\Theta) = \sum_{i=1}^{N} \sum_{c=1}^{C} [-y_{ic} \log(p_{ic})] + \sum_{t=1}^{T} \Omega(g_{tc}), \tag{4.9}$$

where:

- Θ denotes the collection of all model parameters.

This objective function is used to measure the quality of our tree splits and be used to optimise our decision tree.

3. **Using the Objective Function for Tree Building:**

   In XGBoost, each boosting iteration involves constructing a new decision tree designed to reduce the overall objective function. This process begins by computing pseudo-residuals, which are approximated by the first and second derivatives of the loss function with respect to the current

predictions. Specifically, for each instance i, the gradient g_i and Hessian h_i are defined as:

$$g_i = \partial_{\hat{y}^{(t-1)}} \ell(y_i, \hat{y}^{(t-1)}), \quad h_i = \partial^2_{\hat{y}^{(t-1)}} \ell(y_i, \hat{y}^{(t-1)}).$$

These derivatives enable us to form a second-order Taylor expansion of the loss function, leading to an approximate objective for the t-th tree:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t), \tag{4.10}$$

where $\Omega(f_t)$ is the regularisation term as defined earlier.

This approximate objective function is then utilised to evaluate potential splits during tree construction. For any candidate split that divides a node into left (L) and right (R) child nodes, the algorithm computes a gain value, which quantifies the reduction in the objective function resulting from the split. The gain is calculated as:

$$\text{Gain} = \frac{\left(\sum_{i \in L} g_i\right)^2}{\sum_{i \in L} h_i + \lambda} + \frac{\left(\sum_{i \in R} g_i\right)^2}{\sum_{i \in R} h_i + \lambda} - \frac{\left(\sum_{i \in P} g_i\right)^2}{\sum_{i \in P} h_i + \lambda} - \gamma, \tag{4.11}$$

where P denotes the parent node and λ and γ are regularisation parameters. This gain function directly measures the improvement in the objective function resulting from a specific split.

By iteratively selecting the splits that maximise this gain, the algorithm constructs the decision tree in a manner that systematically reduces the overall objective function. Each new tree, added sequentially, concentrates on correcting the residual errors left by preceding trees. In this way, the quality of prediction improves as each subsequent tree is added. This process repeats until a certain set limit of iterations is reached. In this project, we set this limit as 50.

4. **Regularisation Constraints:**

During tree growth, the following constraints were used to help control complexity:

- `max_depth` (4 - 8): This parameter limits the maximum number of splits from root to leaf node. This was optimised via grid search with the `caret` package.
- `gamma` ($\gamma = 0$ - 0.2): This parameter sets a minimum loss reduction criteria for node splitting. Essentially, splits are only accepted when they meet a certain threshold, which is penalised with γ.

5. **Model Update:**

After adding a new tree, predictions are updated by adjusting the contribution of the latest tree using the learning rate η that scales each tree's contribution:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(\mathbf{x}_i). \tag{4.12}$$

Finally, the model computes a score for each class c by aggregating the outputs of an ensemble of T decision trees. For an instance with feature vector $\mathbf{x}_i$, the score is defined as:

$$f_{ic}(\mathbf{x}_i) = \sum_{t=1}^{T} \eta \, g_{tc}(\mathbf{x}_i), \tag{4.13}$$

where:

- $f_{ic}(\mathbf{x}_i)$ is the aggregated score for class c for instance i,
- $g_{tc}(\mathbf{x}_i)$ is the output of the t-th tree for class c,
- η is the learning rate,
- T is the total number of trees.

The predicted probability for class c is then obtained using the softmax function:

$$p_{ic} = \frac{\exp(f_{ic}(\mathbf{x}_i))}{\sum_{k=1}^{C} \exp(f_{ik}(\mathbf{x}_i))}, \tag{4.14}$$

where:

- C is the total number of classes.

This iterative process allows XGBoost to refine its predictions progressively, improving classification performance with each new tree.

4.1.3.3 Evaluation and Limitations

The 1S-XGB model demonstrated inconsistent predictive performance across test seasons, with Elo ratings emerging as dominant predictors. This is different from the 1S-GLM, where xG was the more important metric. As shown in Table 4.3, Away Elo achieved the highest feature importance, closely followed by Home Elo. The higher importance of Away Elo reinforces the earlier findings that the form of the away team was more important in determining the result of the game, which was first suggested in the creation of the 1S-GLM. While xG metrics showed comparatively lower influence, their relative contribution changed gradually across seasons. Away xG moving average importance rose from 0.073 in 2021 to 0.144 in 2023.

Table 4.3: Feature Importance in 1S-XGB Model

Feature	**2021**	**2022**	**2023**
Away Elo	0.427	0.423	0.374
Home Elo	0.411	0.359	0.365
Away xG	0.073	0.115	0.144
Home xG	0.090	0.103	0.116

The model achieved less competitive performance against other modelling approaches, as evidenced in Tables 4.4 to 4.6. The model did not outperform the 1S-GLM generally, and did not consistently outperform some of the other models we have tried. Three key limitations warrant consideration:

First, the model's foundation in historical Elo ratings introduces inherent path dependency. While effective for capturing team quality trends, this approach may undervalue sudden tactical changes or squad developments not reflected in gradual rating adjustments, but better captured in data such as xG.

Second, the gradient-boosted trees' opaque decision processes complicate causal interpretation—though feature importance metrics provide macro-level insights, individual prediction rationales remain obscured.

Finally, the underprediction of draws emerged as a limitation, as we saw earlier in the 1S-GLM, possibly caused by both structural and data-related constraints. As the least frequent outcome, draws suffered from inherent class imbalance, with the model's loss function prioritising accuracy on the more prevalent home and away win classes. This bias manifested in systematically compressed draw probabilities. A possible cause of this issue was that the feature set lacked metrics critical to identifying draw-prone matches. While Elo ratings and xG metrics captured team quality and offensive momentum, they proved insufficient for modelling the equilibrium of closely contested matches. Furthermore, the model's architecture inherently favoured clear-cut predictions, as gradient-boosted trees disproportionately rewarded confident splits between home/away wins rather than probabilistic estimations of balanced outcomes. The underprediction of draws could also have been caused by the absence of explicit draw-focused regularisation or weighted loss components, leaving the minority class vulnerable to marginalisation. Addressing this limitation would require either rebalancing techniques (e.g., synthetic draw oversampling) or modelling adjustments to better capture the distinct features underlying drawn matches.

These limitations notwithstanding, the 1S-XGB framework provides valuable predictive insights while maintaining computational efficiency. Future enhancements could explore multi-stage modelling to handle the modelling of draws separately.

4.2 Two-Stage Models

Addressing the weaknesses of the earlier models, a more novel approach was attempted. Two-stage models were considered, where the different outcomes of the match would be modelled sequentially with separate models. The prob-

abilities of each subsequent outcome were conditioned on the probability of the previous event not occurring. The aim was to allow for a higher degree of flexibility in the modelling process, so variables and modelling techniques that were more effective for predicting draws could be specified, whilst retaining the overall effectiveness of the model.

4.2.1 Two-Stage `glmnet` and XGBoost Model (2S-GLM-XGB)

The model will be referred to as the **Two-Stage `glmnet` and XGBoost Model (2S-GLM-XGB)**. Stage One was modelled with a `glmnet` model, and Stage Two was modelled with a XGBoost model.

1. The probabilities of drawing and not drawing were first modelled with a multinomial logistic repression with a penalised log-likelihood, in a similar fashion to the 1S-GLM.

2. The probabilities of winning or losing were modelled with a xGBoost model, in a similar fashion to the 1S-XGB. The win/loss probabilities were then conditioned on the probability of not drawing which were obtained from the model in the first stage.

The flowchart below describes the modelling process:

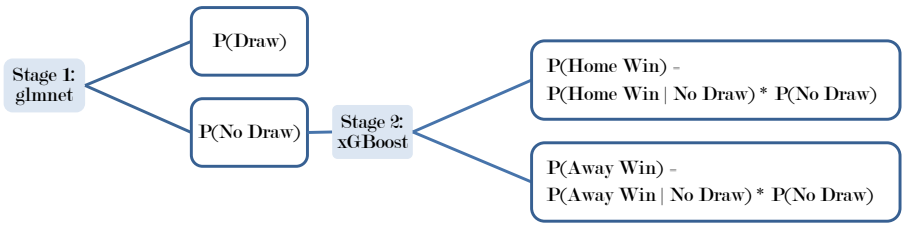

Figure 4.2: Probability Tree Diagram for 2S-GLM-XGB

4.2.1.1 Evaluation and Limitations

The two-stage GLM-XGB architecture failed to deliver meaningful improvements over simpler approaches, constrained by two fundamental limitations.

First, the elastic net regularisation in Stage 1 exhibited unstable feature selection across test years. While the 2021 model employed all predictors with modest coefficients (Home Elo: 1.23×10^{-5}, Away Elo: 4.91×10^{-6}), subsequent iterations nullified critical features through excessive regularisation. The 2022 model retained only Home Elo (0.00019), while the 2023 version retained both Elo impacts (Home: 0.00083, Away: 0.00041) whilst eliminating xG effects entirely. This coefficient instability undermined the model's capacity to maintain consistent prediction logic across seasons. Furthermore, this suggests that the variables present in the model are not very good predictors of the draw outcome.

Second, inconsistent hyperparameter optimisation in Stage 2. The XGBoost stage's optimal configurations varied illogically between test years - learning rates (η) alternated between conservative (0.01 in 2022) and moderate values (0.1 in 2021/2023), while column sampling ratios fluctuated from 0.6 to 1.0. This parameter instability reflects the challenges of tuning decoupled stages through separate cross-validation processes, where local optima in one stage could have created suboptimal conditions for the other.

The model's theoretical flexibility is an advantage that has not been fully utilised. Given more time, it could be possible to better fine-tune the stages of the model to better suit their purposes. However, at this current level, the 2S-GLM-XGB fails to surpass the simpler 1S-GLM, which achieved superior performance.

4.2.2 Two-Stage Random Forest and XGBoost Model (2S-RF-XGB)

The **Two-Stage Random Forest and XGBoost Model (2S-RF-XGB)**. Stage One was modelled with a Random Forest model, and Stage Two was modelled with a XGBoost model.

1. The probabilities of drawing and not drawing were first modelled with a random forest.

2. The probabilities of winning or losing were modelled with a xGBoost model, in a similar fashion to the 1S-XGB. The win/loss probabilities were then conditioned on the probability of not drawing which were obtained from the model in the first stage.

The flowchart below describes the modelling process:

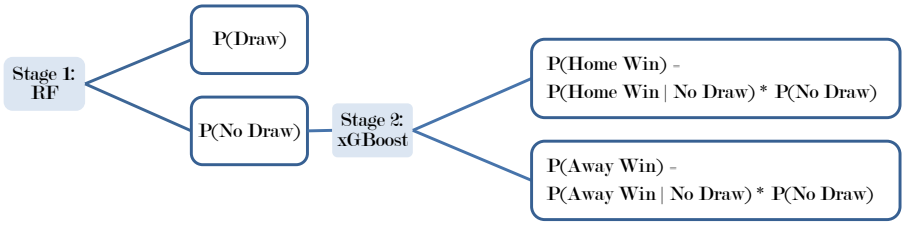

Figure 4.3: Probability Tree Diagram for 2S-RF-XGB

Random Forest is an ensemble learning method that constructs multiple decision trees through bootstrap aggregation. For our binary classification task (Draw vs No Draw), the model estimates probabilities as:

$$P(\text{Draw}|\mathbf{x}_i) = \frac{1}{B}\sum_{b=1}^{B}\mathbb{I}(T_b(\mathbf{x}_i) = \text{Draw}), \tag{4.15}$$

where:

- $B = 500$: Number of trees in the forest (fixed by `caret` defaults)
- $T_b(\mathbf{x}_i)$: Class prediction from b-th tree for observation i
- $\mathbb{I}$: Indicator function returning 1 if tree prediction is Draw

4.2.2.1 Decision Tree Construction

Each decision tree T_b in the ensemble is constructed through three synergistic mechanisms. First, a bootstrap sample comprising approximately 63.2%[1] of the training data is drawn with replacement, preserving the original data distribution while ensuring each tree's uniqueness. Second, feature subsetting introduces diversity by randomly selecting $mtry = 2$ predictors from the four available features—specifically, Home/Away Elo ratings and xG moving averages—at each candidate split. Finally, recursive partitioning employs greedy optimisation through iterative node division, maximising the reduction in Gini impurity:

$$\Delta G(t) = G(t) - \left(\frac{N_{\text{left}}}{N} G(t_{\text{left}}) + \frac{N_{\text{right}}}{N} G(t_{\text{right}}) \right), \tag{4.16}$$

where $G(t) = 1 - \sum_{c \in \{\text{Draw, NoDraw}\}} p_c(t)^2$ quantifies node impurity, N denotes the parent node's observation count, and $N_{\text{left}}/N_{\text{right}}$ represent child node sizes. This combination of stochastic sampling and impurity-driven splitting results in decorrelated trees.

4.2.2.2 Evaluation and Limitations

The 2S-RF-XGB model underperformed relative to comparator models (Tables 4.5-4.6), despite its theoretical capacity to capture non-linear relationships. This discrepancy primarily stems from inherent limitations in probability calibration. Random Forests exhibit poor calibration for minority classes such as draws, which constitute only 20–25% of match outcomes. Stage 1's overconfident draw probability estimates introduced systematic biases that propagated multiplicatively into Stage 2's conditional win/loss predictions.

4.3 Statistical Evaluation of Models

The classification performance of the predictive models is assessed using the following metrics:

[1] Derived from $1 - 1/e \approx 0.632$, representing the expected proportion of unique observations in bootstrap sampling with replacement.

- **Accuracy**: The proportion of correct predictions.
- **Confusion Matrix**: A detailed breakdown of true positives, false positives, true negatives, and false negatives for each class.
- **Logarithmic Score (Log Score)**: Sum of negative log values of the predicted probabilities of actual outcomes.
- **Brier Score**: Mean squared error of predicted probabilities.

The models are compared based on their performances during 3 separate English Premier League seasons, between 2020 to 2023:

Model	**Accuracy (%)**	**Log Score**	**Brier Score**
1S-GLM	50.53%	390	231
1S-XGB	51.05%	398	240
2S-GLM-XGB	51.05%	394	234
2S-RF-XGB	47.89%	398	238
Bookmakers (Max)	**51.58%**	**382**	**227**

Table 4.4: Model Comparison Results (Season 2020/21)

Model	**Accuracy (%)**	**Log Score**	**Brier Score**
1S-GLM	56.58%	364	216
1S-XGB	53.95%	393	236
2S-GLM-XGB	53.16%	391	234
2S-RF-XGB	50.79%	398	233
Bookmakers (Max)	**57.02%**	**327**	**194**

Table 4.5: Model Comparison Results (Season 2021/22)

Model	**Accuracy (%)**	**Log Score**	**Brier Score**
1S-GLM	53.16%	372	222
1S-XGB	52.89%	371	221
2S-GLM-XGB	55.26%	372	221
2S-RF-XGB	51.32%	397	234
Bookmakers (Max)	**56.05%**	**367**	**218**

Table 4.6: Model Comparison Results (Season 2022/23)

For better visualisation, the log scores of the models can be represented in the figure below:

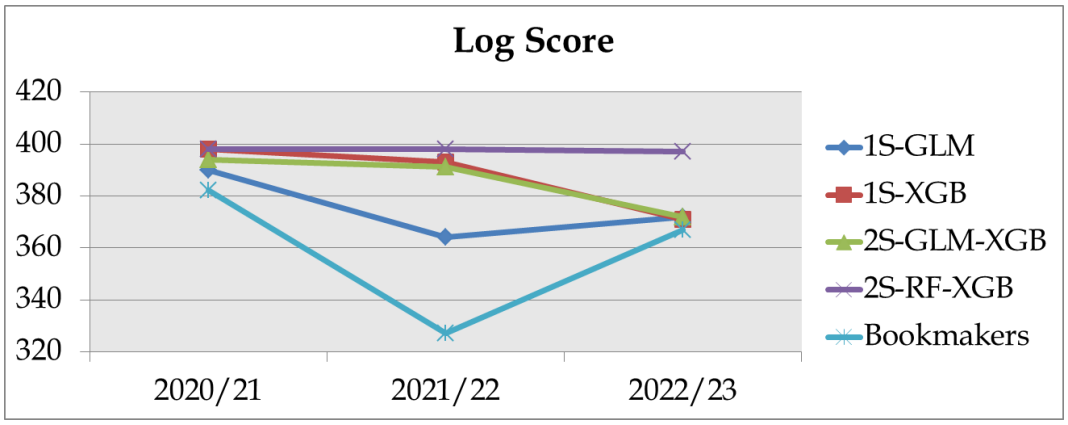

Figure 4.4: Log Scores

4.3.1 Best Model: 1S-GLM and Observations Across Seasons

Most of the models demonstrated similar accuracy, from 47 to 57% accuracy. The 1S-GLM model returned average accuracy ranging from 50.53% in the 2020/21 season to 56.58% in the 2022/23 season. While accuracy scores are generally comparable across all the models, the log scores and Brier scores provided a more meaningful interpretation of the quality of the predictions.

The 1S-GLM model achieved the most competitive scores. The model recorded the lowest log scores in two of the three seasons. Similarly, the model consistently maintained low Brier scores, demonstrating its superior performance in estimating the probability of match outcomes.

Regularisation, a key feature of 1S-GLM, plays a significant role in mitigating overfitting, ensuring that predictions generalise well across different datasets. This suggests that simpler, well-calibrated models are better suited for predicting football outcomes compared to more complex machine learning approaches.

As seen in Figure 4.4, models incorporating the GLM in their architecture consistently exhibited lower log scores compared to their non-GLM counterparts. Across all three seasons, the 1S-GLM and 2S-GLM-XGB models generally outperformed the 1S-XGB and 2S-RF-XGB models, respectively. This trend reinforces that the regularised GLM component contributes positively

to the model's ability to produce more accurate probabilistic predictions.

4.3.2 Mediocre Performance of 2S Models

The two-stage models, 2S-GLM-XGB and 2S-RF-XGB, were designed to combine the strengths of two separate models to better predict the elusive draw. However, their performance was mediocre when evaluated across seasons.

The 2S-GLM-XGB model exhibited reasonably strong performance in the 2022/23 season, achieving a log score of 372 which equalled the 1S-GLM. Its Brier score of 221 likewise remained competitive within the model cohort. However, across the 2020/21 and 2021/22 seasons, the 2S-GLM-XGB model distinctly underperformed relative to the 1S-GLM, with noticeably poorer log scores. This pattern suggests that incorporating a secondary modelling stage not only fails to substantially enhance predictive capability but may impair forecasting accuracy in certain competition contexts. This implies potential redundancy in the model structure or overfitting to training data in specific seasonal contexts. This performance deficiency further indicates that the modelling approach chosen for predicting draws has not been fully optimised. Consequently, further parameter fine-tuning and architectural refinements would be beneficial to address these shortcomings.

Conversely, the 2S-RF-XGB model exhibited consistently inferior performance across all seasons. With log scores of 398, 398, and 397 for the respective seasons, and correspondingly elevated Brier scores, this model substantially underperformed relative to both the alternative models. This marked deficiency could be attributed to the compounding of predictive uncertainties between the random forest model and the XGBoost model. The model's accuracy percentages, ranging from 47.89% to 51.32%, further corroborate this performance gap, suggesting fundamental limitations in the 2S-RF-XGB combination for football match prediction.

Overall, the two-stage modelling framework demonstrates that stacking approaches can yield benefits under specific conditions, but their effectiveness is heavily contingent upon model compatibility and feature diversity. The em-

pirical results across three seasons strongly suggest exercising caution when implementing sophisticated architectures, particularly when simpler, more interpretable models such as the 1S-GLM already produce robust and consistent predictions. This observation follows the principle of parsimony in statistical modelling, where additional complexity should only be introduced when it delivers meaningful performance improvements. The visualisation in Figure 4.4 further illustrates this conclusion, showing that the one stage models tended to outperform the two stage models across multiple competitive seasons.

4.3.3 Beating the Bookmakers

The bookmakers' exhibited inconsistent performance across the three seasons. More particularly, there was a large improvement in the log scores of bookmaker predictions in the 2021/22 season, when compared to the predictions from our models. First, we can analyse the proportion of match outcomes in each of the seasons, as it was earlier found that the models were much poorer at predicting draws.

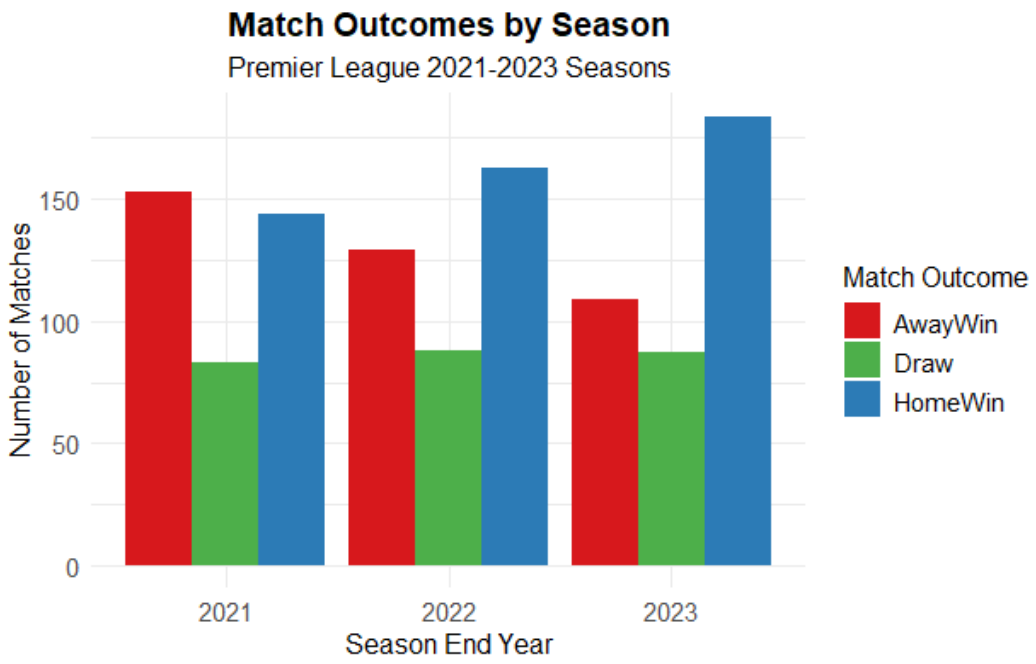

Figure 4.5: Match Outcomes by Season

The proportion of draws in each season remained largely the same. While

the proportion of home wins and away wins varied, there was no clear outlying outcome in 2021/22 which would explain the enhanced scores.

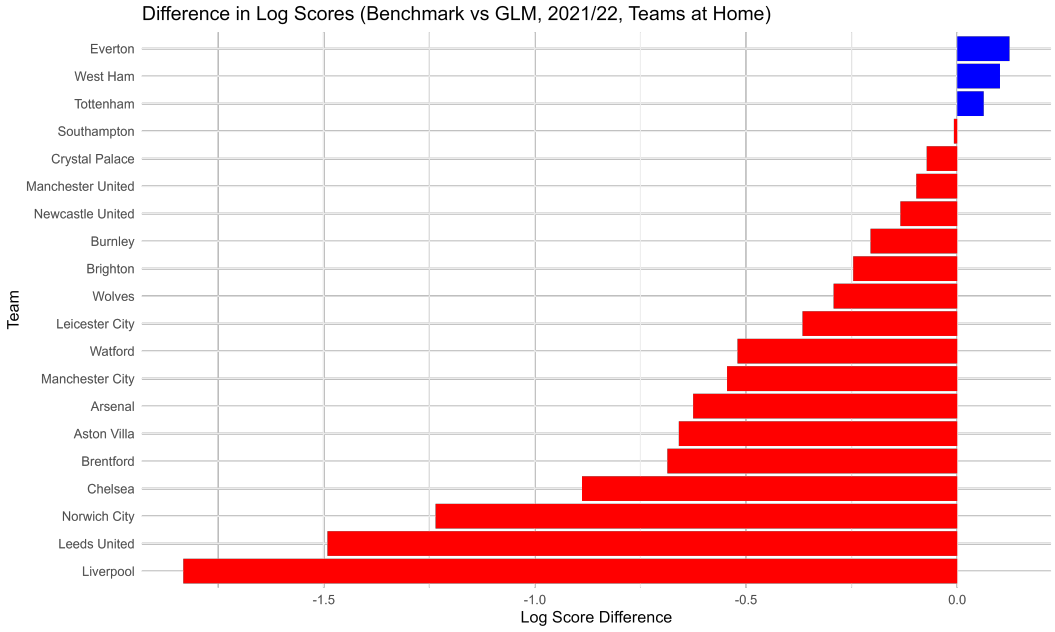

Figure 4.6: Difference in Log Scores (Benchmark vs GLM, 2021/22, Teams at Home

Figure 4.6 shows the difference in log scores between the bookmakers and our best performing model, the 1S-GLM in 2021/22. From this, we see that the biggest difference is in the log scores for Liverpool. One possible explanation for this discrepancy lies in the bookmakers' ability to incorporate real-time information, such as team news, injuries, and public sentiment, which our models do not explicitly account for. For example, during the 2021/22 season, Liverpool was competing for four trophies and maintained an exceptionally high level of consistency in their performances. This dominance likely led to an overwhelming number of bets being placed on Liverpool, forcing bookmakers to adjust their odds to protect their margins. By tightening the odds on Liverpool, bookmakers artificially enhanced their log scores, as the implied probabilities became more aligned with the actual outcomes.

However, that was not always the case, as we see in the following analysis that teams that do poorly can create problems for the 1S-GLM as well.

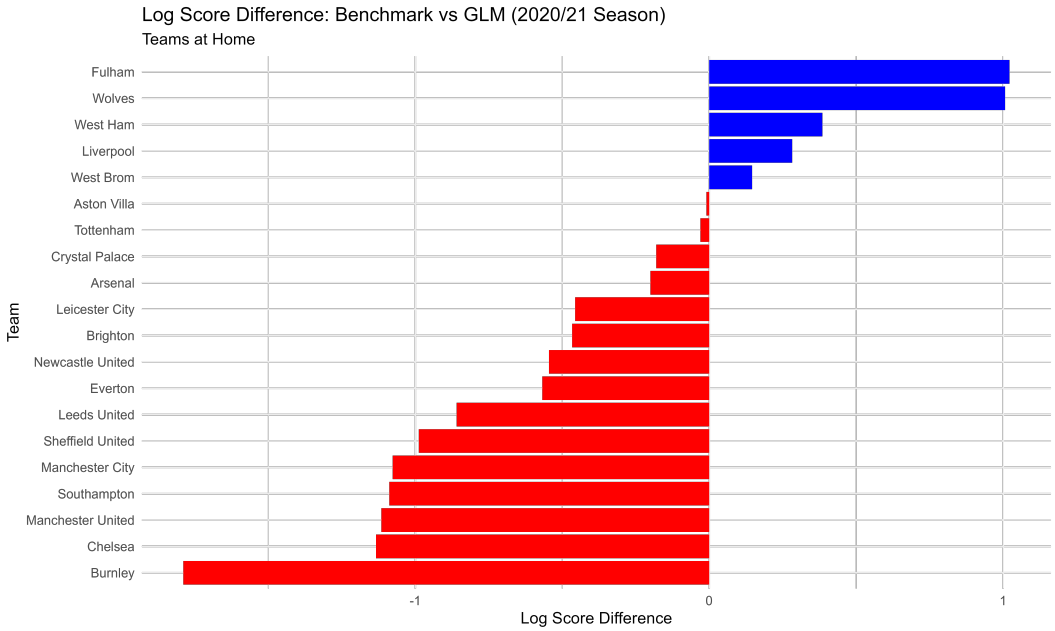

Figure 4.7: Difference in Log Scores (Benchmark vs GLM, 2020/21, Teams at Home

Figure 4.7 shows the difference in log scores between the bookmakers and the 1S-GLM in 2020/21. In this season, we see that the predictions for Burnley under our model had a much poorer average log score than the bookmakers.

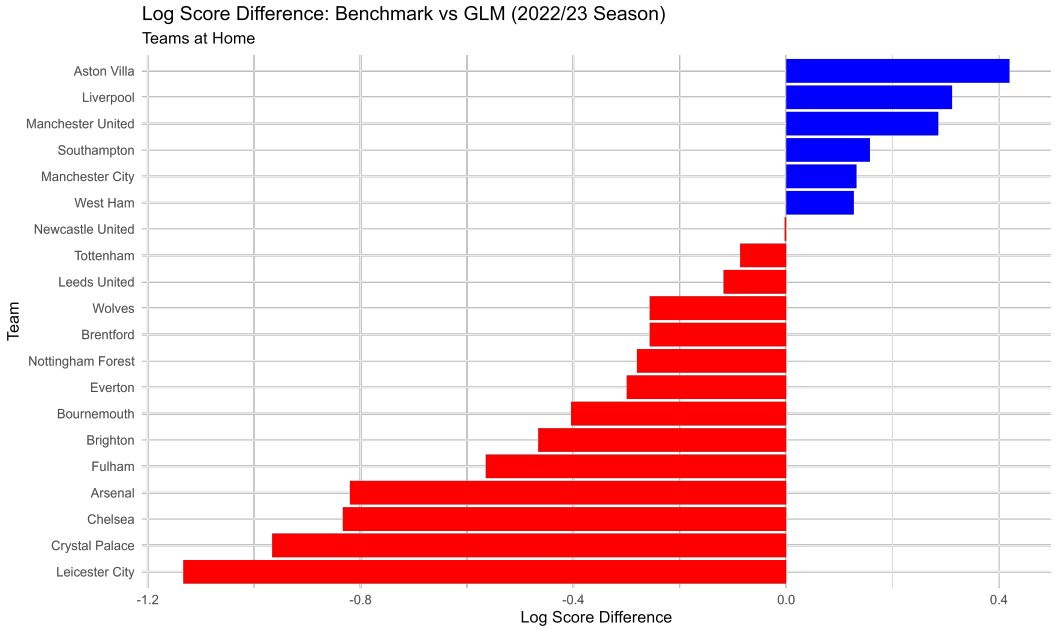

Figure 4.8: Difference in Log Scores (Benchmark vs GLM, 2022/23, Teams at Home

Similarly, in Figure 4.8, which shows the analysis for 2022/23, we see that our model's predictions for Leicester City are much poorer than that of the bookmakers. Interestingly, both Burnley and Leicester City had finished among the bottom four teams in those two seasons, with Leicester City being relegated. Another common feature between the two teams is that they had finished in the mid-table the year before, with Burnley finishing at 10th and Leicester City finishing at 8th the seasons before, respectively. This implies that our models generally struggle when teams experience a drastic shift in form over the summer, whether due to changes in management, key player departures, or tactical overhauls. Going back to our earlier analysis in Liverpool, we see that Liverpool had finished 3rd the season before 2021/22. Going from a top-four team to being title contenders was a large shift which was hard for our model to capture.

This highlights a key weakness of our models: the inability to incorporate real-time sentiments, contextual information, and off-season changes that human analysts and bookmakers intuitively assess. Consequently, in the next section, the formulation of betting strategies must account for these limitations and possibly, even exploit them to our advantage. By identifying scenarios where significant shifts in team dynamics are not fully reflected in bookmaker odds or captured by our models, we can potentially uncover value bets that leverage the discrepancy between statistical predictions and market sentiment.

Chapter 5

Betting Framework

The betting strategies were evaluated against the different models specified earlier. The implied probabilities by the bookmakers was calculated by finding the best odds offered for each outcome across all bookmakers and normalising them. In the chapter, we often consider the Return on Investment (ROI), which is defined as the gain or loss relative to the fixed bankroll of £1,000.

5.1 Bookmaker Strategy

This strategy is to simply bet on the most likely outcome, based on implied probabilities from bookmaker odds, with a fixed stake per game. Given the higher accuracy of the bookmaker's predictions, it would unlikely be more profitable to use the classification presented by our own models. Assuming that the bettor always managed to find the best odds for each bet across the major bookmakers in the UK market (an important assumption, given that all bookmakers have their own margins), the results of this strategy are summarised as follows.

Table 5.1: Baseline Bookmaker Strategy

Season	Number of Bets	Total Bankroll	Total Profit	ROI
2020/21	380	1000	-37.4	-3.74%
2021/22	380	1000	73.4	7.34%
2022/23	380	1000	48.7	4.87%
Total	**1140**	**3000**	**84.7**	**2.82%**

This will serve as a benchmark for the evaluation of the next strategies.

5.2 Basic Edge Strategy

The Basic Edge Strategy is implemented by using the calculated "edge" for each possible outcome of a match (Home Win, Away Win, or Draw). The edge represents the difference between our model's adjusted probability for an outcome and the implied probability derived from the bookmaker's odds. This strategy identifies and places bets based on which outcome has the highest edge, indicating that our model has identified a potential value opportunity.

Despite not being the best statistical model, the 2S-RF-XGB Model performed the best with this strategy.

Table 5.2: Model Edge Strategy, 2S-RF-XGB Model

Season	Number of Bets	Total Bankroll	Total Profit	ROI
2020/21	380	1000	227	22.70%
2021/22	380	1000	1.66	0.166%
2022/23	380	1000	-14.1	-1.41%
Total	**1140**	**3000**	**214.56**	**7.15%**

The high degree of returns from the year 2020/21 season would seem to be an anomaly. This would suggest that the strategy with this model is not sustainable in the long term.

5.3 Negative Edge Strategy

The Negative Edge Strategy is implemented by identifying outcomes where the model's adjusted probability is lower than the implied probability derived from the bookmaker's odds. This indicates that the bookmaker's odds may be overestimating the likelihood of a particular outcome, presenting a potential value opportunity. The strategy places bets on the outcome with the most negative edge, effectively betting against the bookmaker's implied probability.

This strategy is particularly useful in scenarios where the bookmaker's odds are skewed due to public sentiment, recent team performance, or other external factors. By focusing on outcomes with negative edges, the strategy aims to exploit mispriced odds and generate profit over time.

The 1S-GLM model demonstrated the most profit with the Negative Edge Strategy. Its performance is summarised in Table 5.3.

Table 5.3: Negative Edge Strategy, 1S-GLM Model

Season	Number of Bets	Total Bankroll	Total Profit	ROI
2020/21	380	1000	129	12.9%
2021/22	380	1000	59.9	5.99%
2022/23	380	1000	-53.2	-5.32%
Total	**1140**	**3000**	**135**	**4.52%**

It appears that with the exception of the 2022/23 season, the strategy managed to achieve returns. Looking from a more granular perspective, we can look at which outcomes are more profitable under the strategy.

Table 5.4: Negative Edge, Profit by Outcome

Bet	Occurence	Total Profit	Profit/Bet	% Profitable Bets
Home Win	311	0.500	0.00161	32.2%
Draw	425	-58.2	-0.137	12.2%
Away Win	404	193	0.478	22.3%

It is evident that the Away Win bets were the most profitable under the Negative Edge Strategy. This suggests that bookmaker odds may have systematically overestimated the likelihood of home teams winning, leading to value opportunities in backing away teams.

Interestingly, Home Win bets barely broke even, with a negligible profit and a relatively higher percentage of profitable bets (32.2%). Home teams may have been favoured by the bookmakers, leading to suppressed odds, therefore unable to generate significant value.

On the other hand, Draw bets performed the worst, with the lowest percentage of profitable bets (12.2%). This follows the earlier observed difficulty with predicting draws.

Overall, the findings suggest that the Negative Edge Strategy was most effective when applied to away wins, while betting on draws resulted in significant losses. This insight can be used to refine the strategy by selectively

placing bets on away wins where a negative edge is detected, while avoiding draws to minimise losses.

5.4 Sharpe Ratio Strategy

The Sharpe Ratio Strategy optimises bet sizing based on the Sharpe ratio, a measure of risk-adjusted returns. This strategy places greater emphasis on bets with higher expected returns relative to risk, inspired by the strategy adopted by Hubáček et al. (2019).

The Sharpe ratio is calculated as follows:

$$S = \frac{E[R]}{\sigma} \tag{5.1}$$

where:

- S is the Sharpe ratio,
- $E[R]$ is the expected return of the betting strategy,
- σ is the standard deviation of returns.

Bets were placed on the outcome with the highest Sharpe Ratio, when it exceeded 1.5. As this did not apply to all matches, the £1,000 allocated to each season was retrospectively split equally. This allows for ease of comparison between the seasons and strategies. Table 5.5 summarises the performance of the Sharpe Ratio Betting Strategy across multiple seasons. The most profitable model for this strategy was the 1S-XGB Model.

Table 5.5: Sharpe Ratio Strategy, 1S-XGB Model

Season	Number of Bets	Total Bankroll	Total Profit	ROI
2020/21	380	1000	172	17.2%
2021/22	380	1000	-16.4	-1.64%
2022/23	328	1000	37.0	3.70%
Total	**1088**	**3000**	**193**	**6.43%**

In a similar fashion to the Basic Edge Strategy, the Sharpe Ratio strategy demonstrated high profits in the first year. However, this strategy would appear more consistent that the Basic Edge Strategy.

To analyse the strategy from a more granular persepctive, the breakdown of profitability by bet type is provided in Table 5.6.

Table 5.6: Sharpe Ratio, Profit by Outcome

Bet	Occurrence	Total Profit	Profit/Bet	% Profitable Bets
Home Win	319	10.8	0.0340	22.6%
Draw	283	-43.7	-0.154	25.8%
Away Win	486	226	0.465	20.0%

The results indicate that the Sharpe Ratio Strategy is particularly effective for Away Win bets, which contributed the majority of the profits. It suggests that the model is successfully identifying value in away team odds. The relatively lower profitability of Home Win bets suggests that home teams may be overvalued by bookmakers, leading to fewer profitable opportunities. This also aligns with the findings under the Negative Edge Strategy.

Overall, the strategy demonstrates superior profitability compared to others, achieving a more consistent return.

5.5 Bet Selection

As evidenced above, the Away Win is the most profitable match result to bet on. An immediate improvement would be to restrict our bets to Away Wins. In practice, it would be difficult to allocate £1,000 across all the games, as our bet choice is only known on a week to week basis. However, for the sake of comparison, £1,000 will be retrospectively and equally allocated for the games in which the Away Win has been selected as the bet.

In addition, the Negative Edge Strategy and the Sharpe Ratio Strategy will be the strategies to focus on, as they have demonstrated that they are more consistent and profitable.

After restricting the bets to Away Wins, the results are as follows:

Table 5.7: Away Wins Only - Negative Edge Strategy, 1S-GLM Model

Season	Number of Bets	Bankroll	Profit	ROI
2020/21	212	1000	266	26.6%
2021/22	104	1000	186	18.6%
2022/23	88	1000	-26.50	-2.65%
Total	**404**	**3000**	**425.50**	**14.2%**

Table 5.8: Away Wins Only - Sharpe Ratio Strategy, 1S-XGB Model

Season	Number of Bets	Bankroll	Profit	ROI
2020/21	165	1000	567	56.7%
2021/22	178	1000	26.0	2.60%
2022/23	144	1000	-62.2	-6.22%
Total	**487**	**3000**	**531**	**17.7%**

As shown in Tables 5.7 and 5.8, this refinement yields promising results, albeit with some variability across seasons. Under the Negative Edge Strategy using the 1S-GLM model, the restricting the bets to Away Wins delivered positive returns in the first two seasons, with a particularly strong ROI of 26.6% in 2020/21 and 18.6% in 2021/22. However, the 2022/23 season recorded a minor loss of -2.65%, slightly dampening the overall profitability. Nevertheless, across all three seasons, the strategy generated a total profit of £425.50 from a £3,000 bankroll, representing a respectable overall ROI of 14.2%.

The Sharpe Ratio Strategy, paired with the 1S-XGB model, also benefitted from the focus on Away Wins. The 2020/21 season was especially profitable, producing a 56.7% ROI. While the 2021/22 season yielded only modest gains (2.6%), and the 2022/23 season ended in a small loss (-6.22%), the cumulative return still stood at 17.7%. These results suggest that concentrating on Away Wins can be a viable enhancement to both strategies, filtering out lower-value bets and focusing capital on where the models identify the greatest edge. However, the reduced performance in the most recent season serves as a reminder of the volatility inherent in sports betting.

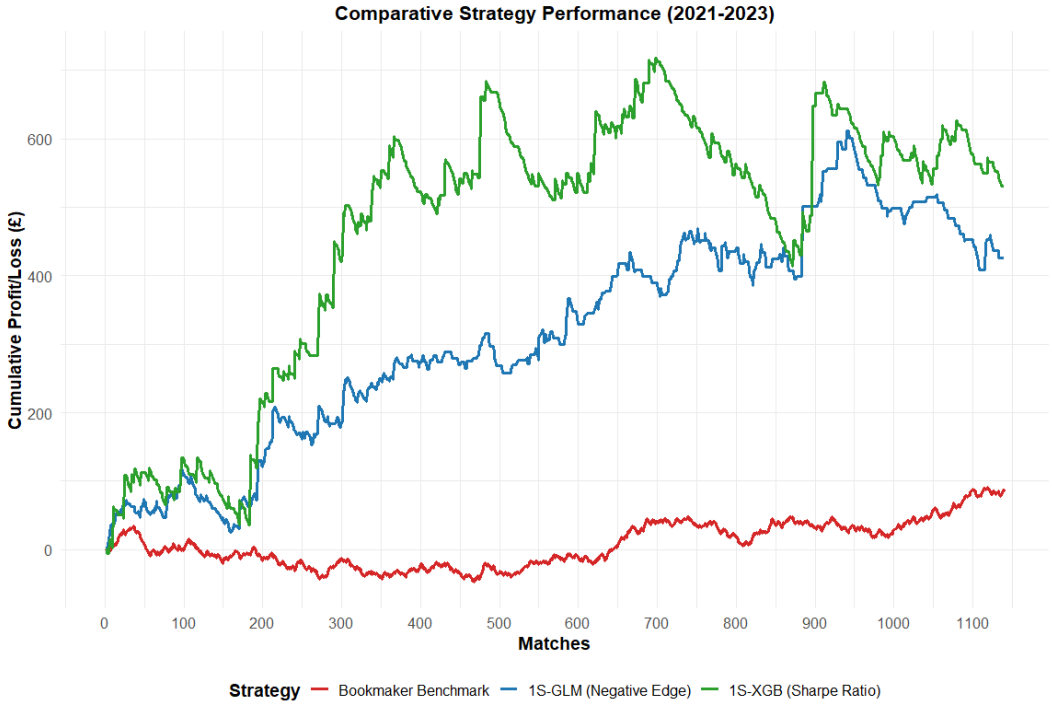

Figure 5.1: Cumulative PnL for Away Win Bets

In examining the performance over time (Figure 5.1), it becomes evident that the Sharpe Ratio Strategy exhibits higher volatility relative to the Negative Edge Strategy. While this heightened volatility can yield higher gains in favourable conditions, it also increases the risk of losing more, as seen in the 2022/23 season. In contrast, the Negative Edge Strategy demonstrates a more stable trajectory, suggesting that it may be a safer long-term strategy.

Interestingly, there are points where the two strategies demonstrate a degree of negative correlation (at Match Numbers 500 and 700). A future idea would be to develop an algorithm to systematically identify periods of negative correlation between the strategies, which could allow for strategy switching (in a manner similar to quantitative trading techniques in financial markets) thereby capitalising on whichever model is performing optimally at any given time.

Looking ahead, there is potential value in the diversification across these strategies. Although both frameworks target Away Wins, their distinct risk-return profiles imply that holding using both may help balance overall portfolio

variance and smooth out cumulative profit-and-loss (PnL). As a result, a hybrid approach could offer a more resilient betting system, combining the consistency of the Negative Edge Strategy and the opportunity for higher returns afforded by the Sharpe Ratio Strategy.

Chapter 6

General Conclusions

6.1 Main Findings

Our analysis reveals three key insights.

First, there is still a predictive performance gap between our statistical models and bookmakers' predictions, as our models consistently underperform in comparison. This discrepancy is likely due to bookmakers' ability to integrate real-time sentiments and contextual information into their odds pricing mechanisms, thereby affecting their implied probability. Future development of our models should aim to include these variables.

Second, the models struggle with draw outcomes, with both the one-stage and two-stage models failing to predict them accurately. Similarly, the betting strategies failed to make a profit when betting on draws. These findings suggest that draws represent an inherently unpredictable component of football matches and might be better omitted from optimised betting portfolios.

Lastly, despite these challenges in predictive accuracy, our project suggests that carefully constructed betting strategies can yield positive returns, particularly by focusing on the Away Win outcome. This confirms that strategic bet selection can, to a certain extent, compensate for imperfect prediction accuracy.

6.2 Future Work

To further enhance predictive performance, future work should incorporate additional variables. For instance, including player ranking or ratings (potentially derived from video game ratings) could capture individual skill differentials more effectively (Fahey-Gilmour et al., 2019). Moreover, integrating injury reports and data on player availability would help account for squad changes, which can have a substantial impact on match outcomes.

Further refinement of the betting strategies could improve returns. Restricting bets to outcomes where value has been consistently demonstrated, such as Away Wins, could improve profitability. Additionally, our strategy placed a fixed stake on all games—introducing a mechanism to dynamically adjust the betting stake based on certain criteria like the Sharpe Ratio might allow for greater capital allocation to high-value opportunities. The cumulative PnL graphs also suggest a degree of correlation between strategies; therefore, identifying quantitative markers could enable the strategic switching of models, or alternatively, the maintenance of both strategies to benefit from diversification.

6.3 Concluding Statement

This investigation establishes that, although outperforming bookmakers' predictive accuracy remains challenging, systematic approaches that utilise dynamic betting strategies show significant promise in generating consistent profits. The framework developed herein offers both a solid methodological foundation and practical insights for future football betting research.

Appendix A

Modelling Parameters

A.1 One-Stage Models

A.1.1 Generalised Linear Model (1S-GLM)

- **Regularisation**: α in a range between 0 to 1, λ between 10^{-2} to 10^{1}
- **Tuning**: 10-fold cross-validation

A.1.2 XGBoost Model (1S-XGB)

- **Hyperparameters**:
  - Max tree depth: 4 - 8
  - Learning rate (η): 0.01 - 0.3
  - Subsample ratio: 0.8
  - Column subsampling: 0.6 - 0.8
- **Stopping Criteria**: 50 boosting rounds

A.2 Two-Stage Models

A.2.1 First Stage (Draw Prediction)

- **Objective**: Binary classification (Draw/NoDraw)
- **Features**: Home/Away Elo ratings, xG moving averages

- **Models Tested**:
  - GLMnet with elastic net regularisation
  - Random Forest (mtry = 5 - 15 via tuning)

A.2.2 Second Stage (Win/Loss Prediction)

- **Objective**: Binary classification (HomeWin/AwayWin)
- **Features**: Home/Away Elo ratings, xG moving averages
- **Models**: XGBoost

Appendix B

Colophon

This document was set in the Modern Latin Roman typeface using LaTeX and BibTeX, composed with the Overleaf text editor.

AI tools were used in an assistive capacity.

Bibliography

Rahul Baboota and Harleen Kaur. Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, 35(2):741–755, 2019. doi: 10.1016/j.ijforecast.2018.12.002.

Adam Bate. Liverpool's former director of research Ian Graham explains how data helped the Reds win the Premier League title. *Sky Sports*, 2024. URL `https://www.skysports.com/football/news/11669/13200552/liverpool-s-former-director-of-research-ian-graham-explains-how-data-helped-the-reds-win-the-premier-league-title`.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.2307/2334029.

Department for Digital, Culture, Media & Sport. Gambling and lotteries: Taking part survey 2019/20, 2020. URL `https://www.gov.uk/government/statistics/taking-part-201920-gambling-and-lotteries/gambling-and-lotteries-taking-part-survey-201920`.

Mark J. Dixon and Stuart G. Coles. 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, 1997. doi: 10.1111/1467-9876.00065.

Arpad E Elo. *The rating of chessplayers, past and present.* Arco Publishing, New York, 1978.

J. Fahey-Gilmour, B. Dawson, P. Peeling, J. Heasman, and B. Rogalski. Multifactorial analysis of factors influencing elite Australian football match outcomes: a machine learning approach. *International Journal of Computer Science in Sport*, 18(3):100–124, 2019. doi: 10.2478/ijcss-2019-0020.

FIFA. Fifa World Cup Qatar 2022 watched by 5 billion people. 2023. URL `https://www.fifa.com/tournaments/mens/worldcup/qatar2022/news/fifa-world-cup-qatar-2022-watched-by-5-billion-people`.

John Goddard. Regression models for forecasting goals and match results in association football. *International Journal of Forecasting*, 21(2):331–340, 2005. doi: 10.1016/j.ijforecast.2004.08.002.

House of Commons. The Gambling Act 2005: A bet worth taking? 2012. URL `https://publications.parliament.uk/pa/cm201213/cmselect/cmcumeds/421/421.pdf`.

Ondřej Hubáček, Gustav Sourek, and Filip Železný. Exploiting sports-betting market using machine learning. In *2019 7th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, pages 93–98. IEEE, 2019. doi: 10.1016/j.ijforecast.2019.01.001.

Lewis and Nabbi. Brentford FC: Premier League club sifts through over 85,000 players using data and 'good eyes'. *CNN*, March 2023. URL `https://edition.cnn.com/2023/03/10/football/brentford-moneyball-success-premier-league-spt-intl/index.html`.

M. J. Maher. Modelling association football scores. *Statistica Neerlandica*, 36 (3):109–118, 1982. doi: 10.1111/j.1467-9574.1982.tb00782.x.

Premier League. The numbers that show this has been a season like no other. 2024. URL `https://www.premierleague.com/news/4031506`.

Havard Rue and Oyvind Salvesen. Prediction and retrospective analysis of soccer matches in a league. *The Statistician*, 49(3):399–418, 2000.

Statista Research Department. Betting industry in the UK - statistics & facts, 2024. URL `https://www.statista.com/topics/6133/betting-industry-in-the-uk/`.

The Footy Tipster. A Brief History of Football Betting, 2023. URL `https://thefootytipster.com/a-brief-history-of-football-betting/`.

Alkeos Tsokos, Santhosh Narayanan, Ioannis Kosmidis, Gianluca Baio, Mihai Cucuringu, Gavin Whitaker, and Franz Király. Modeling outcomes of soccer matches. *Machine Learning*, 108(1):77–95, 2019. doi: 10.1007/s10994-018-5741-1.