Comparing the Predictors: a study of ELO based models for predicting match outcomes in the English Premier League

Emmanuel Odujebe
Mohammad Yusuf Baig
Jesse Olugbire
Felix Jihong Feng
Zongxi Li

Year 4 Project
School of Mathematics
University of Edinburgh
March 7, 2025

Abstract

Elo-based ranking models are widely used to predict team performance in football, but there remains an open question when compared to betting odds for their predictive features. This study evaluates three variants of Elo — Standard Elo ($\mathbf{ELO_b}$), Goal-Difference Adjusted Elo ($\mathbf{ELO_G}$), and Expected Goals-Based Elo ($\mathbf{ELO_{xG}}$) — to determine their effectiveness in forecasting Premier League match results. By using historical match data, we computed team rankings iteratively and converted bookmaker odds into implied probabilities for direct comparison. To assess model accuracy, we used Kendall's rank correlation with final league standings, quadratic loss functions, and Mann-Whitney U tests to measure statistical significance between models. Results showed that while all Elo models correlated strongly with final league rankings ($\tau = 0.768$), betting odds provided more accurate match outcome predictions ($\tau = 0.821$). Moreover, by grouping teams, the results indicated that among three Elo-based models, $\mathbf{ELO_b}$ performed comparably to betting odds for top and bottom teams, while $\mathbf{ELO_{xG}}$ exhibited higher predictive loss across most cases. We concluded that Elo-based models tracked top and bottom teams well as their ratings shifted consistently with results. However, the middle-eight teams were harder to predict due to fluctuating performance. While this makes Elo-based models less responsive to sudden changes, it remains an aide for assessing long-term team performance trends.

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Emmanuel Odujebe*
Mohammad Yusuf Baig
Jesse Olugbire
Felix Jihong Feng
Zongxi Li)

Contents

Chapter 1

Introduction

Rankings and prediction of team performance are fundamental challenges in modern sports analytics, influencing league standings, betting markets, and team management. Among the numerous ranking methods proposed, the Elo rating system, initially developed by Elo (1978), has been widely adapted for team sports such as football(Düring et al., 2022). According to Hvattum and Arntzen (2010), variants of the Elo model incorporate factors such as goal differences, expected goals (xG), and time-dependent adjustments. However, while Elo-based rankings are a widely used measure of a team's relative strength, their predictive accuracy remains an open question when compared to other methods, such as betting odds.

Some studies have explored the applicability of Elo scores in sports prediction. Hvattum and Arntzen (2010) examined Elo-based features in ordered logit regression models and found that they could predict match results competitively against other statistical approaches. Similarly, Düring et al. (2022) extends the Elo to account for changes in team strengths over time, thus demonstrating the adaptability of the system to dynamic team performance. Additionally, betting markets propose an alternative, as odds incorporate expert judgment and market-based probabilities. Some studies have suggested that betting odds can be as good as traditional statistical models due to the continuous refinement of bookmaker(Forrest et al., 2005).

This article aims to evaluate the predictive accuracy of different Elo-based ranking methods, comparing standard Elo, goal-difference-adjusted Elo (ELOg), and expected goals-based Elo (ELOxG) against betting odds. The comparison focuses on the relevance of these ranking systems to actual league rankings and their accuracy in predicting match outcomes based on statistical tests. To achieve this, we analyze:

- The correlation between different ranking models and actual league standing.
- How betting odds compare to Elo-based predictions.
- Statistical significance tests to assess model performance.
- The impact of different Elo update rules, including goal-difference and expected goals adjustments.

- The stability of Elo rankings throughout the season and how initial ratings influence final standings.
- The role of home-field advantage in Elo-based ranking adjustments.
- Variability in the predictive accuracy of betting odds over time.

This project provides an overview and evaluation of how different ranking approaches compare in predicting football matches in Premier League outcomes. The results will help determine whether Elo-based models are sufficient for predictive features, whether refinements such as goal-based adjustments improve their accuracy, and how they compare against betting odds.

Chapter 2

Literature Review

2.1 Introduction

Sports analytics has expanded greatly since the early 2000s, propelled by improved data availability, computational power, and broader acceptance of quantitative methods. As researchers and practitioners seek robust ways to rank teams and forecast matches, they frequently go beyond traditional league tables. Standard standings can be limited: for instance, they focus on aggregate points without explicitly comparing each team's match-by-match strength of schedule or recent form. To address these issues, rating systems that produce dynamic, continuous assessments of team skill have become a major focus in analytics (Elo, 1978; Glickman, 1999; Hvattum & Arntzen, 2010).

Among these rating systems, the Elo rating system, originally developed for chess in the mid-20th century, has gained traction in team sports such as football (soccer), basketball, and hockey. Although Elo was designed for a different purpose (individual matchups in chess tournaments), its iterative update framework has proven flexible for league-based sports (Silver, 2015).
In parallel, academics have explored Poisson-based approaches (e.g., Dixon–Coles models) for modeling match scores (Dixon & Coles, 1997; Karlis & Ntzoufras, 2003) and regression-based methods for outcome prediction.
This review examines these rating and predictive models, emphasizing how they have been adapted to football. It also explores additional factors (e.g., expected goals, betting odds) and discusses the trade-offs in model transparency, predictive accuracy, and other objectives. Finally, the review surveys common evaluation metrics and highlights directions for refining Elo in modern football analytics.

2.2 The Elo Rating System

2.2.1 Historical Foundations of Elo

The Elo rating system was introduced by Arpad Elo for chess (Elo, 1978). Although the earliest versions of Elo date to the mid-1960s, they were formalized in his 1978 monograph. In contrast to many league competitions, chess commonly features multiple tournaments throughout the year, with each participant facing a variety of opponents. Originally, ratings were updated after each tournament, rather than after each individual match, reflecting the structure of competitive chess at the time. Nevertheless, contemporary applications often update ratings at a finer resolution (e.g., after each football match), which is a modern adaptation rather than part of Elo's original design.

Elo's expected outcome formula can take various functional forms, though a commonly cited version uses a logistic or normal cumulative distribution in the rating difference. One standard logistic form is:

$$E_A = \frac{1}{1 + 10^{\frac{R_B - R_A}{400}}}$$

Here, R_A and R_B are the numerical ratings of participants A and B. In classical chess Elo, the difference of 400 in the denominator is somewhat arbitrary and could be replaced by another scaling factor. After a competition, the updated rating is computed by:

$$R'_A = R_A + K(S_A - E_A),$$

where S_A is the actual score (1 for a win, 0.5 for a draw, 0 for a loss), and K is a constant controlling the magnitude of rating adjustments (sometimes referred to as the K-factor).

It should be noted that while the logistic formulation has become standard in modern adaptations, it is not definitively documented that Elo himself derived this function from logistic probability theory. Instead, his original method was more heuristic in nature. Subsequent analyses, such as those by Glickman (1999), have reinforced the utility of a logistic approach for mapping rating differences to winning probabilities.

2.2.2 Adapting Elo to Team Sports

Iterative Updates After Each Match

Modern team-sport adaptations typically apply rating updates after each match, rather than waiting for a tournament to conclude. This procedure has proven useful in league formats, since each team plays multiple matches over a season and points-based standings can be supplemented by a dynamic Elo ranking (Hvattum & Arntzen, 2010). In sports that feature relatively few competitions (e.g., international football, with periodic matches), the iterative Elo approach remains feasible, though the frequency of updates is lower.

Objectives and Trade-Offs in Rating Systems

Rating systems serve multiple objectives that can sometimes conflict:

- **Transparency and Interpretability**: Elo is popular partly due to its conceptual simplicity—a single formula adjusts ratings based on expected vs. actual performance.
- **Predictive Accuracy**: More complex models, such as Poisson-based or machine-learning methods, often yield superior predictive performance. However, they can be less transparent.
- **Stability vs. Responsiveness**: A rating system must balance responsiveness to current form (a higher K-factor) with stability that prevents overreacting to isolated results.

In practice, Elo's relatively transparent mechanics make it appealing for media coverage and fans, although the system can be further tuned or combined with advanced metrics (Silver, 2015).

Rating Changes and the K-factor

The size of an Elo rating change is determined by the product of the difference between the actual score S_A and the expected score E_A, and a constant K:

$$\Delta R = K(S_A - E_A).$$

This means that if a match result is highly expected (i.e., E_A is close to the actual S_A), the rating change is minimal, whereas an unexpected result produces a larger adjustment.

The derivation and calibration of the K-factor involve both empirical analysis and theoretical considerations. One common approach is to use historical match data to optimize K by minimizing prediction error metrics such as mean squared error (MSE) or log loss in forecasting outcomes. For instance, if historical data suggests that a team's performance is volatile, a higher K may be warranted to allow ratings to adapt more quickly. Conversely, for teams with more consistent performance, a lower K can help maintain stability in the ratings.

As an illustrative example, consider a match where a team with an expected win probability $E_A = 0.8$ wins (thus $S_A = 1$). The rating change would be:

$$\Delta R = K(1 - 0.8) = 0.2K.$$

If $K = 20$, the team's rating increases by 4 points; if $K = 30$, the increase is 6 points. Such examples highlight how the choice of K directly affects the sensitivity of the rating system to match outcomes.

Moreover, dynamic adjustments of K based on factors such as match importance, margin of victory, or the inherent uncertainty of a team's initial rating (e.g., for newly promoted teams) can further refine the system. Calibration using cross-validation on historical match data ensures that the selected K and any dynamic adjustments lead to improved predictive performance.

Time Decay and Inactivity Adjustments

A persistent concern in rating systems is whether older matches should be discounted, especially when team compositions or player forms change. Some implementations add time decay, weighting recent matches more heavily or imposing a slight rating decrease when a team is inactive (Rue & Salvesen, 2000). However, standard Elo does not inherently degrade a rating for inactivity. Other rating frameworks (e.g., Glicko or certain ICC cricket ranking systems) make inactivity adjustments to better reflect uncertainty in a team's or player's current strength.

2.3 Alternative Statistical Models

2.3.1 Dixon–Coles Poisson Models

An influential alternative for football analytics is the Dixon–Coles model, which uses Poisson distributions to model each team's goals scored (Dixon & Coles, 1997). In the basic form, the expected number of goals for team i vs. team j is

$$\lambda_i, \quad \lambda_j,$$

with these parameters linked to separate attack and defense strengths for each team, plus a home-field effect. The probability of a particular match score $(X_i = x, X_j = y)$ is then

$$P(X_i = x, X_j = y) \;=\; e^{-\lambda_i}\,\frac{\lambda_i^x}{x!}\;e^{-\lambda_j}\,\frac{\lambda_j^y}{y!},$$

often with a small correlation correction if a low-scoring draw is more or less likely than pure independence would suggest.

Parameters are estimated via maximum likelihood on historical match data (Dixon & Coles, 1997; Karlis & Ntzoufras, 2003). Additional covariates (e.g., injuries, a recent manager change) can be embedded, and time decay can weight recent matches more strongly. Dixon–Coles yields a probability distribution over all possible scorelines, thus capturing richer detail than a pure binary or ternary outcome. However, this approach can be more computationally involved than Elo, and the derived ratings for each team's attack/defense can be less intuitive for non-specialists.

2.3.2 Logistic, Poisson, and Machine-Learning Frameworks

Beyond Dixon–Coles, a range of regression-based or machine-learning methods have been proposed:

- **Logistic Regression**: Predicts the probability of win/loss or multinomial outcomes (win/draw/loss), typically including team indicators, home advantage, form variables, etc. (Ahmed & Campbell, 2012).

- **Generic Poisson Regression**: Models goals for each side independently, akin to Dixon–Coles but without the specialized correlation adjustment (Karlis & Ntzoufras, 2003).

- **Machine Learning (e.g., Random Forests, Gradient Boosting, Neural Networks)**: These can incorporate Elo ratings as one feature among many (like player injury data, recent momentum, or weather) (Hubáček et al., 2019). Some research suggests these ensemble methods can outperform simpler models if given sufficient training data.

Such approaches often produce either direct outcome probabilities or predicted scorelines. Whether they produce stand-alone ratings similar to Elo depends on how the models are structured. Some advanced approaches do yield an interpretable strength parameter for each team, but often the focus is on pure predictive accuracy rather than producing a single ranking index.

To provide additional context and help compare these different methodologies, here is a summary of their key strengths, weaknesses, and ideal use cases.

2.4 Comparative Analysis of Alternative Models

In this section, we summarize each model's key strengths, weaknesses, and ideal use cases.

2.4.1 Elo Rating System

Strengths:

- Simple and transparent.

- Easily updated after each match.

Weaknesses:

- Limited capacity to incorporate contextual factors (e.g., goal margin) unless specifically adapted.

- Assumes fixed parameters unless dynamically adjusted.

Ideal Use Cases:

- Real-time updates in league formats.

- Scenarios where interpretability is crucial.

2.4.2 Dixon–Coles Poisson Models

Strengths:

- Captures full scoreline probabilities.
- Incorporates team-specific attack and defense strengths.

Weaknesses:

- More computationally complex than Elo.
- Requires larger datasets for reliable parameter estimation.

Ideal Use Cases:

- Detailed match outcome predictions.
- Leagues with high scoring variance.

2.4.3 Logistic Regression

Strengths:

- Straightforward implementation.
- Effective for binary or multinomial outcome predictions.

Weaknesses:

- May oversimplify complex match dynamics.

Ideal Use Cases:

- Quick baseline forecasts.
- Integrating additional predictive features with minimal overhead.

2.4.4 Machine Learning Approaches

Strengths:

- High predictive power when sufficient data is available.
- Can integrate diverse features (e.g., injuries, xG, match context).

Weaknesses:

- Less interpretable than simpler models.
- Requires extensive tuning and validation.

Ideal Use Cases:

- Complex scenarios with multiple influencing factors.
- Situations prioritizing predictive accuracy over transparency.

2.5 Incorporating Additional Performance Indicators

2.5.1 Goals and Expected Goals (xG)

Classical Elo updates rely solely on the match result ($S_A - E_A$). However, margin-of-victory adjustments partially address the issue that a 5–0 thrashing is more telling than a narrow 1–0. A further refinement uses expected goals (xG), a metric that estimates the likelihood of scoring for each shot based on factors such as distance, angle, defensive pressure, and shot context (Lucey et al., 2014). By using xG or xG-based margins, an Elo-like system can reflect how dominant a performance truly was, rather than only the final scoreline. This can reduce noise from "lucky" finishes or fluke goals, potentially improving the rating's predictive power.

2.5.2 Betting Odds

Betting odds aggregate market intelligence, capturing unstructured information about injuries, transfers, management changes, or even insider knowledge (Forrest et al., 2005). Comparisons of model predictions to implied probabilities from odds markets often serve as a benchmark for evaluating performance. If an Elo system or a Poisson-based model can match or exceed the accuracy of odds-derived predictions (assessed by proper scoring rules), it suggests the model has effectively captured or surpassed market insights.

2.6 Evaluating Models: Objectives and Scoring Rules

2.6.1 Predictive Accuracy and Correlation with Final Standings

When the objective is to rank teams over an entire season, researchers compare predicted vs. actual league standings. Common metrics include:

- **Mean Squared Error (MSE)**: on points or final rank positions.
- **Correlation Coefficients** (Pearson's r, Spearman's ρ): measures the association between predicted and actual ranks.
- **Kendall's τ**: measures the concordance between two rank orderings, robust to small data fluctuations or ties (Hvattum & Arntzen, 2010).

It is important to clarify whether models predict pre-season expectations for final standings or an evolving forecast updated weekly. An Elo-based system typically aims to capture current team strength at each match day; as such, it can be tested both in real time or post hoc over a season.

2.6.2 Probabilistic Calibration: Proper Scoring Rules

Many models output match-level win/draw/loss probabilities, prompting the use of proper scoring rules to evaluate calibration and sharpness:

- **Log Loss (Cross-Entropy)**: penalizes overconfident incorrect predictions heavily.
- **Brier Score**: sums squared differences between predicted probabilities and actual outcomes.

A proper scoring rule is one that is minimized by stating true underlying probabilities. This ensures that models are rewarded for honesty in their probability estimates, rather than simply picking the most likely outcome (Ahmed & Campbell, 2012). Researchers also examine whether a model outperforms baseline benchmarks, such as a naive home-win probability or the implied probabilities from betting odds.

2.7 Refinements and Research Directions

2.7.1 Deeper Insights into the K-Factor and Decay

While a fixed K-factor is common, research suggests:

1. **Dynamic K**: adjusting K based on rating differences or match importance can better capture upsets or high-stakes games (Hvattum & Arntzen, 2010; Silver, 2015).
2. **Time-Decay Weighting**: scaling historical matches so that recent outcomes influence ratings more. In principle, Elo updates are chronological, but adopting an explicit decay can speed convergence to current team strength (Rue & Salvesen, 2000).
3. **Inactivity Penalties**: decreasing a rating if a team is inactive for an extended period is sometimes employed outside football (e.g., cricket or tennis ranking systems).

2.7.2 Hybrid Approaches

Several studies incorporate Elo as a feature in machine-learning or Bayesian hierarchical models (Hubáček et al., 2019). In these setups, Elo captures "general team strength" while more specialized variables (injuries, scheduling congestion, or advanced metrics like xG) refine predictions at the match level. A Bayesian approach can embed Elo-like rating updates at one level of the hierarchy, with top-level priors adjusting for broader uncertainties (e.g., mid-season roster overhauls).

2.7.3 Integration of Advanced Metrics: Expected Points and Scoring Rules

Some recent advances in football analytics have focused on integrating additional performance metrics to refine Elo-based updates. A prominent example is the use of expected points= (xP), which are derived from advanced shot models or probabilistic frameworks like the Dixon–Coles model. Rather than using fixed outcomes of 1 (win), 0.5 (draw), or 0 (loss), an xP-based approach credits teams with a continuous measure reflecting the expected share of points based on the quality and quantity of scoring opportunities. For instance, if a model estimates that a team had a 70% chance of winning, the team might be awarded an xP value closer to 0.7 rather than a full point, thereby reducing the impact of variance from random events.

In parallel, proper scoring rules such as the log loss and the Brier score are employed to evaluate the calibration and sharpness of these probabilistic forecasts. These scoring rules are designed to be minimized when the predicted probabilities match the true underlying probabilities. For example, if a model predicts a 75% chance of victory for a match and the team wins, the log loss would be computed as:

$$\text{Log Loss} = -\ln(0.75),$$

penalizing overconfident mispredictions. Similarly, the Brier score is calculated as the mean squared difference between the predicted probability and the actual outcome, providing a straightforward measure of forecast accuracy.

By integrating xP into an Elo-based system, the rating updates can more accurately reflect team performance by mitigating the noise inherent in binary outcomes. Moreover, employing scoring rules ensures that the predictive models are well-calibrated and that the probabilities produced are reliable. In practice, the combination of these approaches can lead to a more robust and responsive rating system that adapts quickly to changes in team form and match dynamics.

2.8 Conclusion

The Elo rating system, despite its origins in mid-20th-century chess, has been widely adopted and adapted for modern team sports, including football. While Elo's original mechanics involved updating ratings after tournaments, contemporary applications adjust ratings after each match in a league-based format. Such adaptations reflect the pragmatic need for continuous, real-time assessments of team strength.

Trade-offs abound in rating-system design: Elo is notably transparent and easy to understand, yet more complex frameworks (e.g., Dixon–Coles or ML-based) can yield superior predictive performance. Moreover, the objectives of a rating system (public-facing transparency vs. purely predictive accuracy) can differ, influencing how updates and parameters are chosen. Further refinements—time decay, margin-of-victory weighting, or integration with expected goals—offer paths to improved calibration, while maintaining Elo's intuitive core.

Beyond Elo, Poisson-based and regression-driven models provide deeper insight into the distribution of goals and the underlying factors of match outcomes. Still, hybrid solutions often combine Elo ratings with additional performance indicators (e.g., xG, betting odds) to build ensemble methods that balance interpretability and predictive power. The continued evolution of these approaches underlines the dynamic nature of football analytics, in which rating systems must adapt to changing team conditions, data availability, and the multifaceted demands of stakeholders.

Chapter 3

Methodology

In this section, we discuss the methodology used to create, test and evaluate our Elo-based predictive models. First, we describe the data for training and testing. Second, we briefly describe a basic Elo rating system, and two additional Elo systems using advanced football metrics. Next, we describe the prediction method based on our Elo systems and the benchmark models used to assess these Elo methods. Finally, we discuss the scoring methods used to evaluate model strength after simulating over our data.

3.1 Data

For the purpose of this paper, we use data from five seasons of the English Premier League, going chronologically from the 2019/20 season, and on till the 2023/24 season. This data includes results from each match played in the league, including final scores and expected goals (xG). We also use pre-match betting odds from six different companies, for each match in the 2023/24 Premier League season.

3.2 Elo Rating Systems

We use a basic Elo system with three match results, similar to the one described above. Here, E_H describes the expected rating which is scored by the home team in a match, based on the ratings of the home (H) and away (A) teams, R_H and R_A, respectively:

$$E_H = \frac{1}{1 + 10^{(R_A - R_H)/400}},$$

and

$$E_A = 1 - E_H = \frac{1}{1 + 10^{(R_H - R_A)/400}},$$

where E_A is the expected rating scored by the away team. The updated rating for the home team after the match is then

$$R'_H = R_H + K \times (S_H - E_H),$$

where S_H is the actual result for the home team. Similarly for the away team, we have

$$R'_A = R_A + K \times (S_A - E_A).$$

For the purpose of this paper, we use a rating update magnitude of $K = 32$ for the basic Elo function.

In addition to the basic Elo system, we also use two modified Elo systems: one which incorporates the goal differences in a match result, and the other, xG differences. For the goal difference method, the value of the rating update magnitude, K, is replaced by the expression

$$K = K_0(1 + \delta)^{\lambda},$$

where δ is the absolute goal difference, and K_0 and λ are positive constants. Here, we have used the values $K_0 = 10$, and $\lambda = 0.5$. The above expression for K clearly shows that the larger the goal difference λ between two teams, the more the Elo rating changes, for each match. A team winning a match by a margin of 5 goals leads to a much larger gain in Elo rating than a win by only a single goal.

The xG system for Elo works similarly to the goal difference method. As previously described, teams in each match accumulate a certain amount of xG, with the metric being being used as an attempt to better understand how well a team plays throughout the match. The larger the value of xG for a team, the more high quality goal scoring chances they are said to have had. In this system, a team ″wins″ or ″loses″ Elo after a match purely based on whether they generated more or less xG than the other team, regardless of the actual number of goals scored. Furthermore, we use the same expression for K as used previously with the goal difference method, with the exception of the value of δ being the absolute **xG difference**, rather than the previously used goal difference between the teams.

3.3 Elo Predictive Model

For each of the three Elo systems described in section 3.2, we use the same method for predicting the results. The initial Elo for each team is set to 1500. When a team is promoted to the Premier League, their initial Elo is set to 1400, to reflect their weaker standing when joining the league. Results from each of the individual matches of the first two seasons in the dataset, 2019/20 and 2020/21, are used to provide Elo ratings for each team.

The next two seasons, 2021/22 and 2022/23, are used to train logistic regression models to produce probabilities for each match result: models for home wins and away wins are built independently. The model uses Elo differences between teams ($R_H - R_A$) in each match to generate these probabilities. The probability of a draw result is assigned such that all probabilities sum to 1.

Finally, these probabilities are used to predict the results of the final season, 2023/24. Given the regular points system in association football, the expected points for each team is given by:

$$\mathrm{xP} = 3P_H + P_D,$$

where P_H and P_D are the probabilities of a home win and draw respectively. Final league standings are also calculated by summing the expected points for each team over the course of the season.

For the remainder of this paper, the basic Elo rating system described in section 3.2 will be referred to as $\mathbf{ELO_b}$, whilst the rating systems based on goals and expected goals will be referred to as $\mathbf{ELO_G}$ and $\mathbf{ELO_{xG}}$ respectively.

3.4 Benchmark Models

In order to assess the accuracy of the three Elo models described in section 3.3, we use the pre-match betting odds mentioned in section 3.1 to predict the same season, 2023/24. We use two slightly differing methods of prediction: one using the average odds of each outcome across the six companies (referred to as **AVG** in future sections), and the other using the maximum value of the odds of each outcome (referred to as **MAX** in future sections), from all the companies. For both methods, the odds for each match are converted into match probabilities:

$$P_H = \frac{\frac{1}{O_H}}{\frac{1}{O_H} + \frac{1}{O_A} + \frac{1}{O_D}},$$

where P_H is the probability of a home win, and O_H, O_A and O_D are the pre-match odds for a home win, away win and draw, respectively. The probabilities for a draw and an away win are similarly calculated, and, as previously, these three values are normalised, and will also add up to 1. Finally, these probabilities are used to simulate matches and construct a league table in a similar manner to that used in section 3.3.

3.5 Model Evaluation Methods

We test the accuracy of the results provided by these sets of predictors in two ways. The first is to compare the final predicted league tables with the actual 2023/24 league table. We do this by calculating the Kendall rank correlation coefficient for each predictor. This method quantifies the number of concordant and discordant pairs between the actual league table, and each of the predictive tables. The resulting τ values indicate the similarity between each of the ranking tables when compared to the actual table, enabling us to measure the accuracy of each forecast over the course of the entire season.

However, in order to accurately gauge the effectiveness of each of the predictors, it would be favourable to assess the models on a game-by-game basis. For this reason, the second evaluation method uses the quadratic loss function. The quadratic loss for each match is calculated using the equation

$$L = (P_H - y_H)^2 + (P_A - y_A)^2,$$

where L is the total quadratic loss for a match, P_H and P_A are the actual points won by the home team and the away team respectively (with values 0, 1 or 3, depending on the outcome), and y_H and y_A are the expected points for the match, for the prediction system being assessed. We use the above for each of the predictors, and then calculate the means and standard deviations of the losses for each. Finally, we compare the results for each of the predictors against one another, using statistical tests. Since the losses did not appear to be normally distributed (even when transformed logarithmically), Mann-Whitney U tests are instead used to compare the distributions of results.

Chapter 4

Results

4.1 ELO

As mentioned in section 3.3, the first two season of data, the 2019/20 and 2020/21 seasons, were used to initialise Elo ratings for Premier League teams. An example of how Elo ratings can change over time is shown in Figure 4.1, using $\mathbf{ELO_b}$ as a focus.

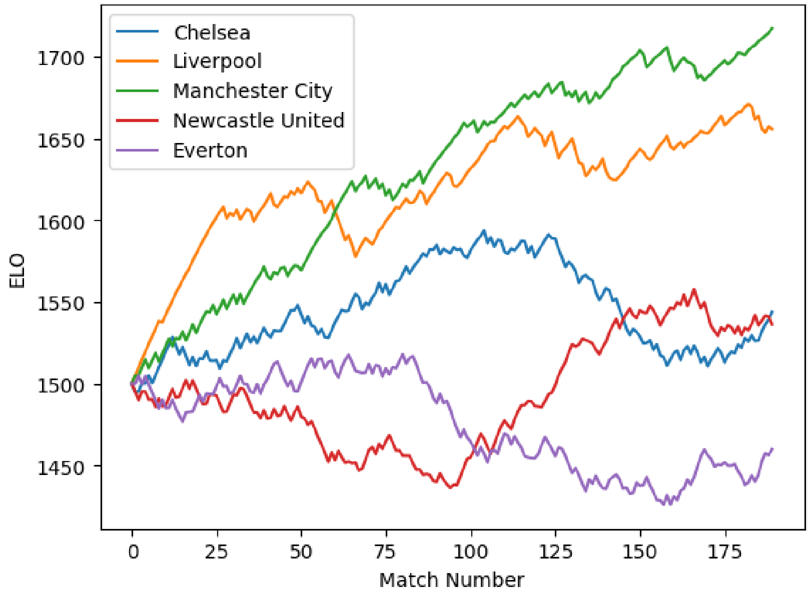

Figure 4.1: Changes in $\mathbf{ELO_b}$ ratings for five Premier League clubs across five seasons.

Manchester City and Liverpool have both had incredibly successful seasons in recent years, with the two clubs claiming the five most recent titles. This is especially the case with Manchester City, who won four in a row. This was reflected in their Elo rating, which steadily increased over time as the club became the highest-ranked team in the league.

Chelsea and Newcastle United have had a more varied time in the Premier

League. Chelsea's final league placements in the five season of data used were 4th, 4th, 3rd, 12th and 6th. These two lower placed finishes were reflected negatively by the club's lower ratings in later seasons. Conversely, Newcastle's league finishes were 13th, 12th, 11th, 4th and 10th. The 2022/23 season's high placement led to a steep increase in the club's Elo rating, although this plateaued in the following season.

A relationship between league placement and Elo rating can be established by inspection. In the case of Everton, the club's progressively worse league finishes (12th, 10th, 16th, 17th, 15th) were shown by declining Elo ratings.

The changes in Elo ratings over time were distinctly more volatile for the $\mathbf{ELO_G}$ and $\mathbf{ELO_{xG}}$ rating systems.

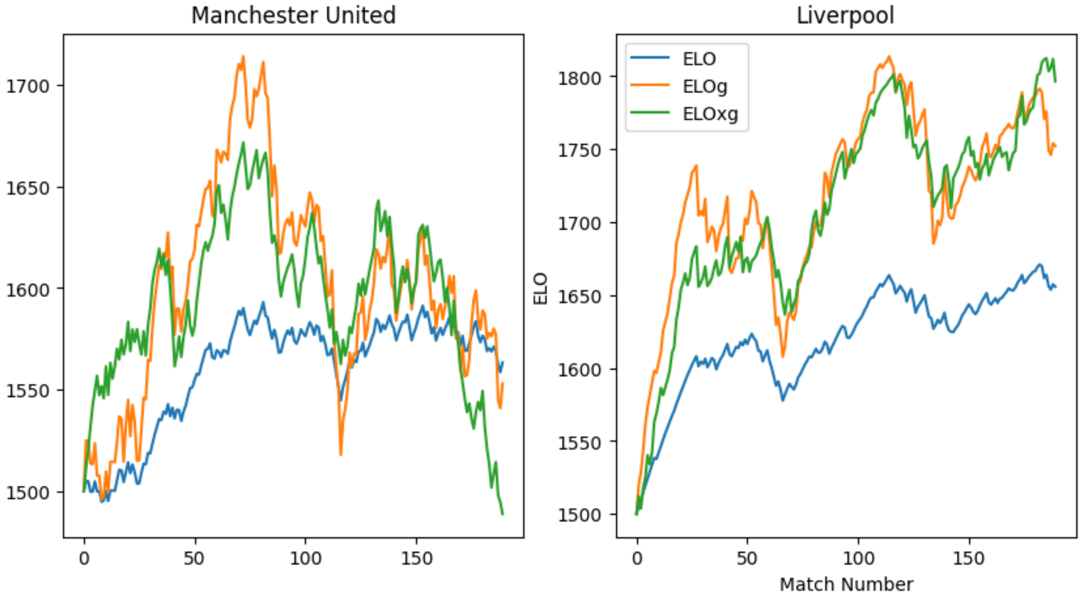

Figure 4.2: Changes in $\mathbf{ELO_b}$, $\mathbf{ELO_G}$ and $\mathbf{ELO_{xG}}$ for Manchester United and Liverpool.

Returning to Liverpool, Figure 4.2 shows the increase in all three Elo ratings over the course of the five seasons. However, ratings based on goals and xG were much higher, due to those ratings granting greater rewards for bigger wins, whilst the results-based $\mathbf{ELO_b}$ ignored the magnitude of victories. This is an indication that, whilst Liverpool did well in regards to pure results, they did even better in regards to goal difference and expected goal difference. For example, in the 2022/23 season, Liverpool finished in 5th place with 67 points. However, their final goal difference of +28 was the 4th highest in the league.

In the case of Manchester United however, the team's poor performances on goals and expected goals were reflected in lower goal-based ratings. This was especially the case with $\mathbf{ELO_{xG}}$. In the 2023/24 season, Manchester United finished in 8th place with 60 points. Their goal difference of -1 also placed them in joint 8th place. However, their xG difference -12.5 was 15th in the league, and was reflected by a sharp decline in the club's $\mathbf{ELO_{xG}}$ rating over the latter matches. This suggested that Manchester United often got ”lucky” with their results, scoring more goals and conceding fewer goals than would be expected given the quality of the chances created by and against them. This was ignored within Manchester United's $\mathbf{ELO_{b}}$ and $\mathbf{ELO_{G}}$ ratings, which did not change much over the same period.

An example of how the magnitude of victory or defeat affected each Elo rating system differently comes in the form of Liverpool's 7-0 victory over Manchester United on March 5th 2023. $\mathbf{ELO_{b}}$ only took into account Liverpool's victory. Based on both teams' Elo ratings at the time, Liverpool gained 4.327 points. $\mathbf{ELO_{G}}$ took into account the 7-goal margin of victory and awarded Liverpool 32.158 points. Despite the large scoreline, the xG scoreline was different, 3.44 xG - 0.84 xG in Liverpool's favour. This 2.6 xG difference was taken into account by $\mathbf{ELO_{xG}}$, which awarded Liverpool 13.518 points.

The highest ratings for the three Elo systems were held by Manchester City: 1717.477 $\mathbf{ELO_{b}}$ and 1885.954 $\mathbf{ELO_{G}}$ (vs West Ham (H), 19 May 2024); 1920.112 $\mathbf{ELO_{xG}}$ (vs Liverpool (A), 16 October 2022).

Similarly, the lowest ratings were all held by Sheffield United: 1356.639 $\mathbf{ELO_{b}}$ and 1221.188 $\mathbf{ELO_{G}}$ (vs Tottenham (H), 19 May 2024); 1214.840 $\mathbf{ELO_{xG}}$ (vs Brentford (H), 9 December 2023).

A full list of final Elo ratings and rankings at the end of the 2023/24 can be found in the appendix.

4.2 Regression Models

As described in section 3.3, two logistic regression models were trained to predict the probability of a home win and an away win for a given match, using games from the 2021/22 and 2022/23 Premier League seasons. Figure 4.3 shows those probabilities as a function of the difference in $\mathbf{ELO_{b}}$ rating (with respect to the home side).

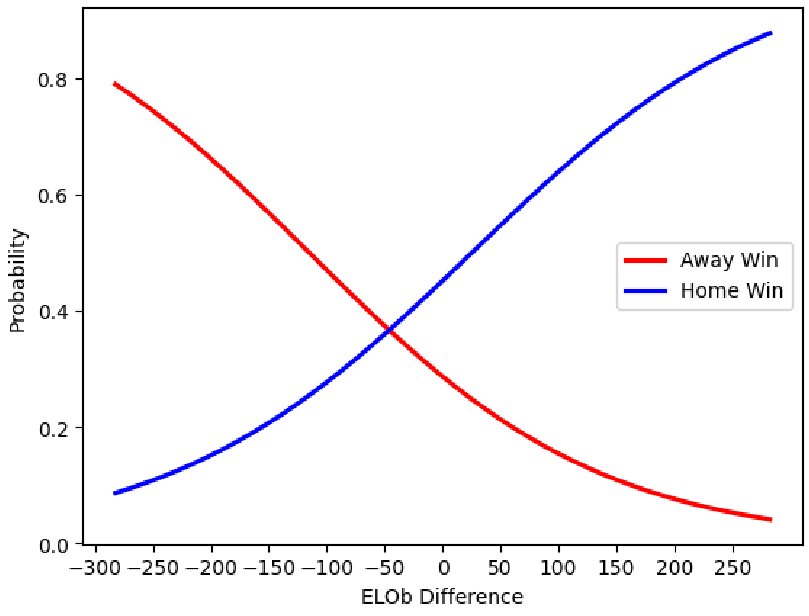

Figure 4.3: Logit probabilities of home and away wins, as a function of difference in $\mathbf{ELO_b}$.

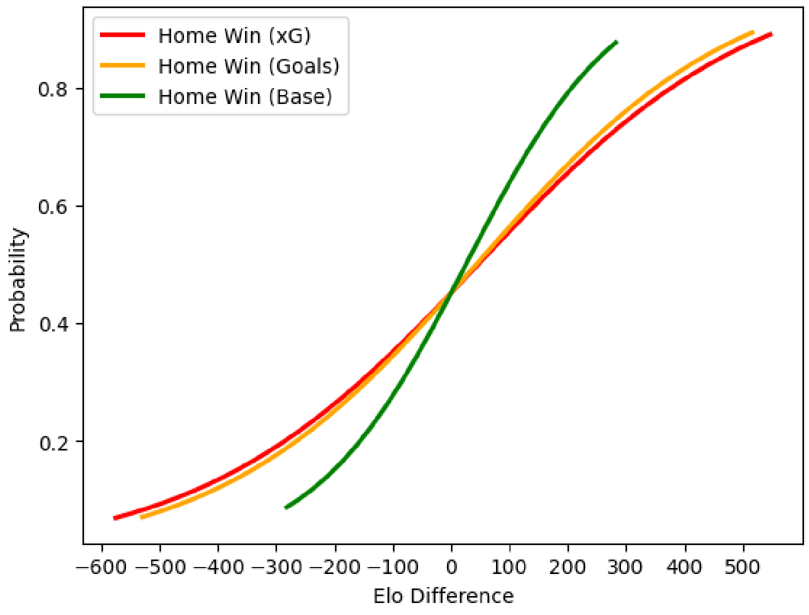

Figure 4.4: Logit probabilities of home wins, as a function of differences in $\mathbf{ELO_b}$, $\mathbf{ELO_G}$ and $\mathbf{ELO_{xG}}$.

The plot of both functions provides an indication of the home advantage in terms of Elo differences. In the case of $\mathbf{ELO_b}$, the probability of winning was equal for both teams when the home team was around 50 points weaker, suggesting that the average home team advantage was about 50 $\mathbf{ELO_b}$. For $\mathbf{ELO_G}$ and $\mathbf{ELO_{xG}}$, the average home team advantage was around 75 points.

Figure 4.4 shows the probabilities of home wins for the three Elo models. Both goal-based models' win probabilities were much less sensitive to changes in Elo difference. Win probabilities were slightly more responsive to $\mathbf{ELO_G}$ than to $\mathbf{ELO_{xG}}$. This suggests that the result of a game by itself has more impact on the results of future games than the magnitude of that result, or the underlying performance.

4.2.1 Home vs Away Results

Given that home and away predictions were modelled separately, as mentioned in section 3.3, it is worth evaluating the performance of each model individually. Table 4.1 shows the coefficients of each model, as well as multiple performance measures, in order of accuracy.

Model	Intercept	Slope	Accuracy	B. Accuracy	Precision	Recall
$\mathbf{ELO_b}$ (Away)	-0.902	-0.008	0.717	0.600	0.602	0.286
$\mathbf{ELO_{xG}}$ (Away)	-0.924	-0.004	0.691	0.514	0.588	0.042
$\mathbf{ELO_G}$ (Away)	-0.921	-0.005	0.689	0.518	0.538	0.059
$\mathbf{ELO_{xG}}$ (Home)	-0.193	0.004	0.653	0.639	0.665	0.481
$\mathbf{ELO_b}$ (Home)	-0.193	0.008	0.651	0.643	0.637	0.550
$\mathbf{ELO_G}$ (Home)	-0.192	0.005	0.650	0.637	0.656	0.490

Table 4.1: Coefficients (intercept and slope), accuracy, balanced accuracy (B. Accuracy) , precision and recall of the home and away variants of the $\mathbf{ELO_b}$, $\mathbf{ELO_G}$ and $\mathbf{ELO_{xG}}$ logistic regression models.

In terms of accuracy, the away models appeared to be best. However, it is worth looking at balanced accuracy as well, of which the away models performed worse than the home models. This is largely due to unbalanced data. Within the training data, away teams won 31% of the 760 matches played. With much fewer away wins present in the data compared to losses and draws, the away models were naturally more conservative in predicting away wins. This was also reflected by the substantially lower recall figures for the away models.

Figures 4.5 and 4.6 show confusion matrices for the home and away variants of the $\mathbf{ELO_b}$ model. The away model in particular had a high number of false positive predictions compared to true positive predictions. This may have been due to the model failing to capture the home advantage shown in section 4.2.

Home advantage is a concept widely accepted and documented concept (Pollard, 2008). It is very often the case that a team perceived to be stronger plays a weaker team away from home and fails to win. This concept was not captured very well within these regression models. Further research and work surrounding the implementation of home advantage could be done to improve

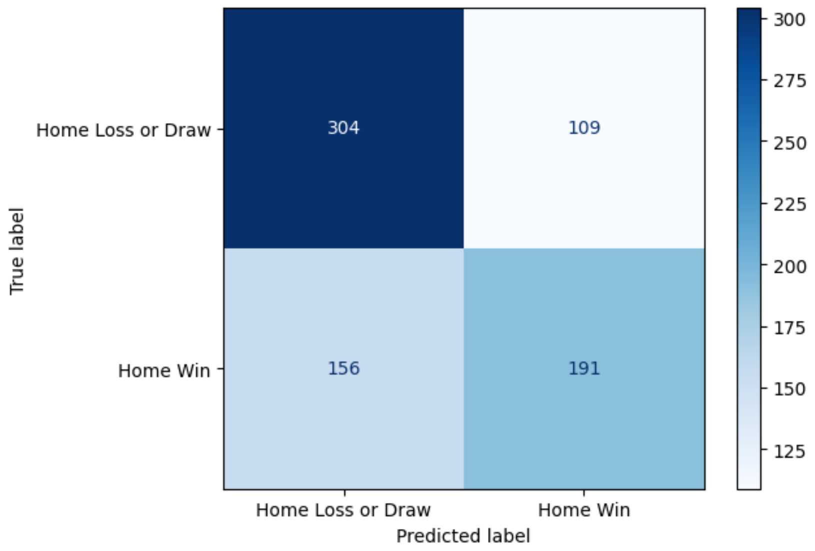

Figure 4.5: Confusion Matrix for the home variant of the $\mathbf{ELO_b}$ logistic regression model.

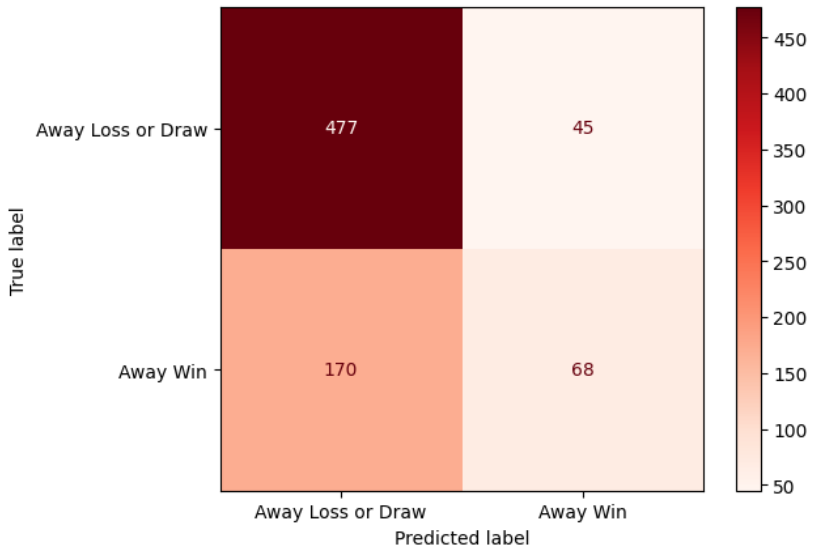

Figure 4.6: Confusion Matrix for the away variant of the $\mathbf{ELO_b}$ logistic regression model.

the performance of Elo-based models.

The home variant of the model had a similar issue, with false negatives often failing to capture the home advantage of weaker teams. However, with home wins accounting for 46% of outcomes, this variant didn't suffer the same issues that the away variant did in regards to unbalanced data. Further work and research on adjusting models to account for unbalanced data would help to create stronger Elo-based models and report the full capabilities of these models within prediction.

4.3 Model Evaluation

As mentioned in section 3.3, the trained Elo-based models were run over the 2023/24 Premier League season, as well as predictive models based on match day odds for each match. Expected points were then calculated for each game, and a final league table was established for each model.

Table 4.2 shows the Kendall rank coefficient associated with each model's final standings when compared to the actual league table at the end of the season. Table 4.2 also shows the average and standard deviation of the total quadratic loss of expected points against actual points for both home and away teams associated with each game, for each model. Table 4.2 finally shows the *p*-value associated with Mann-Whitney U tests for differences in the distribution of average quadratic loss for each model against the average odds-based predictive model (**AVG**).

Model	Rank Coefficient (τ)	Avg Loss	Std	p-value
$\mathbf{ELO_b}$	0.768	2.734	2.709	0.794
$\mathbf{ELO_G}$	0.768	2.866	2.210	0.007
$\mathbf{ELO_{xG}}$	0.768	2.888	2.176	0.003
MAX	0.821	2.638	2.625	0.828
AVG	0.821	2.646	2.581	NA

Table 4.2: Kendall's rank coefficients, mean (Avg) and standard deviation (Std) of total quadratic loss, and *p*-values of Mann-Whitney U tests (against **AVG**) for differences in distribution of quadratic loss for each model.

In regards to the final league table standings, all three Elo-based prediction models had the same performance, with predicted rankings showing a strong correlation ($\tau = 0.768$) to the true final rankings. However, this correlation was weaker than that of the rankings formed by both odds-based prediction models ($\tau = 0.821$). Whilst this suggests that match day odds are more effective in the long run at predicting match results compared to differences in Elo, the strong correlation of Elo-based models provides credibility for the use of Elo in match predictions over the course of a season or seasons.

	$\mathbf{ELO_b}$	$\mathbf{ELO_G}$	$\mathbf{ELO_{xG}}$	**MAX**
Top 6	0.996	0.001	0.000	0.771
Bottom 6	0.760	0.008	0.004	0.835
Middle 8	0.127	0.853	0.059	0.886

Table 4.3: *p*-values of quadratic loss-focused Mann-Whitney U tests about the results of teams ranked in the top six, bottom six and middle eight of the final 2023/24 Premier League season, comparing each model to **AVG**.

Analysis on quadratic loss, unlike rank coefficients, can be done on a match-by-match basis. The higher average quadratic loss of the Elo-based models compared to the odds-based models again highlighted the relative weakness of Elo as a determinant for match outcome prediction compared to match day odds.

Performing a Mann-Whitney U test on the distribution of quadratic loss per match, however, suggested that the difference in average loss between the **AVG** model and $\mathbf{ELO_b}$ model was not significant at any reasonable level. At the 1% level however, the test suggested that the average loss associated with both $\mathbf{ELO_G}$ and $\mathbf{ELO_{xG}}$ was significantly different to **AVG**. This result provides a large amount of credibility towards the potential use of the base Elo rating system in predicting individual matches. There was no significant difference in the distribution of the losses associated with both odds-based models.

4.3.1 Grouped Results

One particular area of interest is the comparative performance of each model on specific groups of results or teams. A simple way of grouping teams is by final league position. Typically, Premier League teams can be split into a ”top six”, ”middle eight” and ”bottom six”. Top six teams are usually competing for European competition qualification places; bottom six teams are usually competing to avoid relegation from the league; middle eight teams are somewhere in between. With different primary objectives and often different levels of strength, it is useful to analyse the performance of these models across these groups. Table 4.3 shows the *p*-values of Mann-Whitney U tests concerning the distribution of quadratic loss for teams that finished within the top six, bottom six and middle eight of the 2023/24 Premier League season, compared to that of **AVG**.

For top six and middle eight teams, both the $\mathbf{ELO_b}$ and **MAX** models showed no significant difference to that of the **AVG** model. The $\mathbf{ELO_G}$ and $\mathbf{ELO_{xG}}$ models exhibited significantly greater loss at all levels, however. For bottom six teams, the $\mathbf{ELO_b}$, $\mathbf{ELO_G}$ and **MAX** performed to the same level as the $\mathbf{ELO_{AVG}}$ model at the 10% level, whilst the $\mathbf{ELO_{xG}}$ showed a significant dif-

	$\mathbf{ELO_b}$	$\mathbf{ELO_G}$	$\mathbf{ELO_{xG}}$	**MAX**	**AVG**
Top 6 vs Bottom 6	0.127	0.284	0.294	0.179	0.189
Top 6 vs Middle 8	0.000	0.023	0.033	0.000	0.000
Bottom 6 vs Middle 8	0.048	0.285	0.335	0.017	0.019

Table 4.4: p-values of pairwise Mann-Whitney U tests for differences in quadratic loss between clubs ranked in the top six, bottom six and middle eight of the final 2023/24 Premier League standings.

ference above the 5% level.

These results suggest that whilst the $\mathbf{ELO_b}$ model performed significantly worse when predicting games involving middle eight teams, it kept in line with the odds-based predictive models across all teams, highlighting its predictive strength.

The $\mathbf{ELO_{xG}}$ model performed significantly worse compared to the odds-based models in all groups, but least-so within the middle eight group. Across these results, the $\mathbf{ELO_{xG}}$ model did not show the same level of predictive strength as $\mathbf{ELO_b}$ or the odds-based models. This suggests that good use cases of $\mathbf{ELO_{xG}}$ may lie outside of prediction. Whilst the model has less use for predictors, a club may find use in the model in keeping track of the team's performances on the pitch. A team with a higher $\mathbf{ELO_{xG}}$ rating in comparison to their $\mathbf{ELO_b}$ rating are likely stronger than their results suggest, due to ”unlucky” shooting across games. This type of analysis can aide in evaluating manager performance: a club may decide to stick with a manager under whom the team has increased in $\mathbf{ELO_{xG}}$ ratings despite poor results and league positions.

Table 4.4 shows p-values for Mann-Whitney U tests that tested for a significant difference in the results between the three groups for each model. For $\mathbf{ELO_b}$ and the odds-based models, there was a significant difference between the average quadratic loss when dealing with middle eight teams compared to other groups of teams. For teams in the middle of the league standings, with an even spread of wins and losses, it followed that Elo ratings did not increase or decrease massively over the course of the season. As such, it was harder to estimate how strong those teams were based on results alone, and $\mathbf{ELO_b}$ became a weaker predictive tool.

Referring back to Table 4.3, the $\mathbf{ELO_G}$ model kept up with the **AVG** model especially well for middle eight clubs, with no significant difference in quadratic loss. This result suggests that for those teams with a more even spread of results, the magnitude of wins and losses may be a better determinant of future results.

Manchester United

As briefly mentioned in section 4.1, the basic results-based Elo rating system did not accurately reflect the decline in performances of Manchester United over the course of the 2023/24 season. This interaction led to a few unique results regarding the $\mathbf{ELO_b}$ model's predictions of Manchester United matches:

- Of the 84 matches that $\mathbf{ELO_b}$ predicted incorrectly, 11% involved Manchester United. Of those games, all but one involved the model predicting a Manchester United win, but the team losing the fixture.
- Of the 10 matches that $\mathbf{ELO_b}$ predicted correctly and **AVG** predicted incorrectly, five of those games involved Manchester United. In all but one of these games, **AVG** predicted a Manchester United loss, and $\mathbf{ELO_b}$ predicted a win.

Whilst the former result suggests a need for time-decay implementation within the $\mathbf{ELO_b}$ model to capture recent form within predictions, the latter result suggests the opposite. Over a long period of time, the basic Elo rating system becomes akin to a numerical representation of a club's reputation, which is slowly built over time and slowly lost over time with consistently good or poor results. This means that losses are often seen by fans as more unexpected for a club as highly reputable as Manchester United, and that unlikely victories (with respect to match day odds) can occur when the team plays up to the standards that their reputation suggests, irrespective of current form.

One can conclude from this case that whilst Elo-based methods do not outperform match day odds in terms of predictive power, there is scope for the use of Elo rating systems as a means of quantifying the long-term growth and decline of a club's results ($\mathbf{ELO_b}$) or on-field performances ($\mathbf{ELO_G}$ and $\mathbf{ELO_{xG}}$), as well as being an aide in wider discussion about a club's over-performance or under-performance against their perceived reputation.

Chapter 5

Discussion

5.1 Conclusion

In this study, we have touched upon the ability to predict association football with the use of three Elo models: first, the base $\mathbf{ELO_b}$; second, the goal-difference-based $\mathbf{ELO_g}$; and finally, its expected value counterpart, the xG-based $\mathbf{ELO_{xG}}$. For each system, we have used past results to initialise Elo ratings, and to train a logistic regression model. We have then used this model to predict matches for the 2023/24 English Premier League season. Parallel to this, we have used pre-match odds — meticulously calculated by several notable betting companies — to predict the same season. The two systems, **AVG** and **MAX**, have been used to benchmark the Elo-based models mentioned prior. Finally, we have evaluated the performance of these models with two separate approaches: the first being the performance of each model over the course of the entire season, and the second being a game-by-game evaluation. In both cases, the results of the model have been compared to the actual match results from the 2023/24 season.

The results of our investigation have shown that over the course of the season, all three Elo predictors have been equally accurate to one another in producing the final league table. And, while they have shown strong correlation to the actual 2023/24 result, they have been slightly outperformed by both the odds-based models. On a game-by-game basis, however, the $\mathbf{ELO_b}$ model has performed to a level that is not significantly different to the **AVG** model. On the other hand, the $\mathbf{ELO_g}$ and $\mathbf{ELO_{xG}}$ models have proven to perform significantly worse than the odds-based models. The three Elo models also appear to predict home results notably better than away results, likely due to the unbalanced nature of the data, as it contains significantly more home wins than away wins.

5.2 Remarks

It is fair to remark that the Elo-based models used in this study are not yet able to outperform pre-match odds for predictive purposes. As mentioned in 2.5.2, betting companies use a large amount of features in the calculation of each set of odds. As such, these results, particularly those achieved by $\mathbf{ELO_b}$, do in fact show a promising start to the prospect of using Elo in predicting association football outcomes.

Naturally, the aforementioned shortcomings of the Elo models can be addressed, leading to enhanced accuracy. This can be achieved through further investigation in expanding upon the following areas:

1. **Initial Elo Assignment:** As mentioned in 3.3, returning Premier League teams are assigned an Elo rating of 1500, while promoted teams at the start of any season are set to 1400. The latter in particular is a result using only Premiership data. A much better way to gauge teams being promoted would be to implement datasets from lower leagues such as the Championship, to more accurately rate promoted teams with a larger sample size of data.

2. **Dataset Size:** Along with the use of lower leagues' data as mentioned in the previous remark, the use of data for previous years would enhance the validity of the test. Using multiple seasons' data for the training of the logistic regression models could improve the accuracy levels. The larger test sample size would also more conclusively assess the Elo models when compared to the benchmarks.

3. **Home Advantage:** The implementation of home advantage modifications to the Elo systems could address the accuracy concerns regarding away results. These could also be addressed by implementing countermeasures towards the unbalanced nature of the data towards away wins, where methods such as oversampling could be looked into as potential solutions, possibly leading to a better representation of home advantage in the away win logistic model.

4. **Predictive Model:** This study made use of logistic regression to formulate the predictors for the Elo-based models. Further research could explore various different machine learning approaches to compare and identify the most accurate predictor for association football results, as mentioned in 2.4.4.

5. **Parameter Variation:** This study used specific parameter values, such as the value of the Elo rating update magnitude, K, for the $\mathbf{ELO_b}$ model, as well as the values of K_0 and λ, for the dynamic K models $\mathbf{ELO_g}$ and $\mathbf{ELO_{xG}}$. For further study, a range of possible values can be tested for each of these parameters, possibly improving accuracy across the board,

as well as potentially bringing the most out of the $\mathbf{ELO_g}$ and $\mathbf{ELO_{xG}}$ models, which underperformed in this study.

6. **Time Decay:** The Elo models used in this study weigh results from the first season equally to the final season. Given the possibility of increasing the size of the training dataset to incorporate more years' worth of data, the introduction of a time decay variable could be explored, as mentioned in 2.2.2. This could potentially lead to improved predictions for recently underachieving teams, as evidenced by the models' failure to accurately portray the decline of Manchester United in this study.

Bibliography

Ahmed, N., & Campbell, M. (2012). On estimating simple probabilistic discriminative models with subclasses. *Expert Systems with Applications*, *39(7)*. https://doi.org/10.1016/j.eswa.2011.12.042

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)*. https://doi.org/10.1111/1467-9876.00065

Düring, B., Fischer, M., & Wolfram, M.-T. (2022). An elo-type rating model for players and teams of variable strength. *Philosophical transactions of the Royal Society of London*, *380(2224)*. https://doi.org/10.1098/rsta.2021.0155

Elo, A. E. (1978). *The rating of chessplayers, past and present*. ARCO PUBLISHING. INC. New York.

Forrest, D., Goddard, J., & Simmon, R. (2005). Odds-setters as forecasters: The case of english football. *International Journal of Forecasting*, *213*. https://doi.org/10.1016/j.ijforecast.2005.03.003

Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *48(3)*, 377–394. https://doi.org/10.1111/1467-9876.00159

Hubáček, O., Šourek, G., & Železný, F. (2019). Exploiting sports-betting market using machine learning. *International Journal of Forecasting*, *35(2)*. https://doi.org/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)*. https://doi.org/10.1016/j.ijforecast.2009.10.002

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(3)*. https://doi.org/10.1111/1467-9884.00366

Lucey, P., Bialkowski, A., Monfort, M., Carr, P., & Matthews, I. (2014). Quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. *Proceedings of the 8th Annual MIT Sloan Sports Analytics Conference*.

Pollard, R. (2008). Home advantage in football: A current review of an unsolved puzzle. *The Open Sports Sciences Journal*, *1*. https://doi.org/10.2174/1875399X00801010012

Rue, H., & Salvesen, Ø. (2000). Prediction and retrospective analysis of soccer matches in a league. *The Statistician*, *49(3)*. https://www.jstor.org/stable/2681065

Silver, N. (2015). How we calculate nba elo ratings. *Five Thirty Eight*. https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/

Appendix A

Additional Tables

Squad	Points	**ELO$_{b}$**	**ELO$_{G}$**	**ELO$_{xG}$**	**MAX**	**AVG**
Manchester City	91	88.751	75.809	74.535	88.525	87.406
Arsenal	89	76.295	67.037	66.181	81.121	80.440
Liverpool	82	81.923	70.886	69.837	78.910	78.248
Aston Villa	68	58.551	55.780	55.537	58.432	58.260
Tottenham	66	64.632	59.541	59.088	62.146	61.770
Chelsea	63	57.108	54.865	54.673	64.049	63.641
Newcastle	60	61.005	57.274	56.948	60.994	60.786
Manchester Utd	60	67.335	61.261	60.714	57.994	57.839
West Ham	52	49.714	50.330	50.396	46.535	46.665
Crystal Palace	49	45.105	47.465	47.6893	43.999	44.095
Brighton	48	55.540	53.915	53.778	56.383	56.232
Everton	48	42.110	45.626	45.954	46.738	46.846
Bournemouth	48	41.132	44.964	45.326	45.528	45.608
Fulham	47	39.814	44.144	44.553	44.501	44.660
Wolves	46	45.865	47.910	48.109	41.943	42.109
Brentford	39	46.499	48.345	48.521	47.849	48.032
Nottingham Forest	36	35.029	41.093	41.667	39.699	40.001
Luton	26	31.839	38.978	39.659	29.943	30.548
Burnley	24	36.387	41.981	42.508	34.961	35.357
Sheffield Utd	16	32.158	39.175	39.846	25.516	26.250

Table A.1: Points; Elo-based and Odds-based expected points for the 2023/24 Premier League season.

Squad	Points	**ELO$_{\mathbf{b}}$**	**ELO$_{\mathbf{G}}$**	**ELO$_{\mathbf{xG}}$**	**MAX**	**AVG**
Manchester City	1	1	1	1	1	1
Arsenal	2	3	3	3	2	2
Liverpool	3	2	2	2	3	3
Aston Villa	4	7	7	7	7	7
Tottenham	5	5	5	5	5	5
Chelsea	6	8	8	8	4	4
Newcastle	7	6	6	6	6	6
Manchester Utd	8	4	4	4	8	8
West Ham	9	10	10	10	12	12
Crystal Palace	10	13	13	13	15	15
Brighton	11	9	9	9	9	9
Everton	12	14	14	14	11	11
Bournemouth	13	15	15	15	13	13
Fulham	14	16	16	16	14	14
Wolves	15	12	12	12	16	16
Brentford	16	11	11	11	10	10
Nottingham Forest	17	18	18	18	17	17
Luton	18	20	20	20	19	19
Burnley	19	17	17	17	18	18
Sheffield Utd	20	19	19	19	20	20

Table A.2: Final rankings for points, Elo-based and odds-based expected points at the end of the 2023/24 Premier League season.