ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

From Ratings to Results: Advanced Predictive Modeling for Soccer Outcomes

Bachelor Thesis

Gilles Vogt

September 5, 2023

Supervisors: Prof. Dr. Ulrik Brandes, Gordana Marmulla
Department of Computer Science, ETH Zürich

Abstract

This research introduces a refined model for soccer match outcome prediction by assigning dynamic ratings to football teams. The model builds upon and extends the approach presented by Constantinou (2019) [2].

A key extension is the incorporation of league ratings, enabling the model to predict outcomes for not only domestic but also international fixtures.

Data from diverse football leagues and competitions serve as the foundation for the model's predictions and undergo comprehensive preprocessing.

An optimization technique is applied to determine the optimal parameters for the rating system.

Subsequently, a logistic regression model utilizes the ratings to predict match outcomes.

The research also introduces a novel rating initialization method for teams transitioning between leagues accounting for the relative strengths of leagues within nations.

The practical utility of the model is demonstrated through a detailed worked example.

The model's performance is then assessed using the Ranked Probability Score [7].

Contents

Chapter 1

Introduction

Association football, commonly referred to as soccer, is the world's most popular ball game, both in terms of participants and spectators [15]. According to the FIFA Big Count survey conducted in 2006 there were already approximately 265 million soccer players globally at that time [8]. It's essential to note that this figure only accounts for those actively playing the sport. The number of fans and enthusiasts, who passionately follow and support soccer, far exceeds this. And considering the sport's steady growth over the years, these figures are likely even higher today.

This widespread interest translates into significant economic value, with the European soccer market alone valued at €29.5 billion during the 2021/22 season [4].

Hence, it is understandable that predictive analytics in soccer has attracted significant attention due to its potential benefits for teams, businesses, and fans. However, accurate prediction remains challenging due to the sport's inherent unpredictability and the multitude of influencing factors, from team dynamics to external conditions. While top-tier leagues possess abundant data, many others face challenges arising from limited historical data, particularly when direct matchups between teams are infrequent.

The Dolores model, as presented by Constantinou [2], offers a data-minimalistic approach to this challenge, providing predictions even in the absence of direct historical data between contesting teams. This innovative approach prompted us to investigate its effectiveness for predicting the outcomes of international games, which often suffer from a lack of extensive historical data, making this a key area of our work.

In this research, the Dolores model served as a foundational reference. However, we opted for logistic regression as an alternative to the original Bayesian Network approach. The primary motivation for this choice was the simplicity and transparency of logistic regression.

Additionally, our work focused on enhancing the rating initialization method for teams transitioning between leagues, recognizing that relegations and promotions require specialized consideration in predictive models. A significant contribution of this research is its capability to predict outcomes for international games. Given the global significance of events like the UEFA Champions League, the ability to provide reliable predictions for such matches is of high interest.

In summary, this thesis critically evaluates the Dolores rating system, introduces methodological enhancements, and expands its application to international soccer games. The findings and discussions presented herein aim to contribute to the broader discourse on predictive methodologies in soccer.

Chapter 2

Soccer Match Outcome Prediction Methodologies

Soccer match outcome prediction has been an area of interest for decades. Over time, methodologies have evolved to leverage available data, computational power, and statistical tools. Here, we broadly categorize the methodologies into three techniques: deterministic models, probabilistic models, and machine learning models.

2.1 Deterministic Models

Deterministic models operate under the premise that the same set of inputs will always yield the same outputs. In the context of soccer predictions, these models rely on predefined formulas or algorithms to determine outcomes, without factoring in inherent randomness.

Notable Implementations:

- *Elo rating system* (1978): Originally designed for chess, it has been adapted for soccer. It modifies team ratings after each match, taking into account the rating difference between the two competing teams [6].
- The *pi-rating system*, introduced by Constantinou and Fenton (2013), serves as the foundation for the Dolores rating system. In our research, we build upon and extend the capabilities of this system [3].
- Since 2018, the *FIFA World Ranking* adopted a variation of the Elo system. This ranking system is used by FIFA to evaluate the relative strengths of national teams and plays a role in seeding for international tournaments [9].

While deterministic models provide specific outcomes, the inherent unpredictability of soccer matches led researchers to explore models that can

capture this randomness.

2.2 Probabilistic Models

Probabilistic models embrace the inherent unpredictability of soccer matches. Instead of aiming for absolute predictions, they focus on estimating the likelihood of various outcomes using probability distributions, capturing the randomness and uncertainties inherent in the sport.

Notable Implementations:

- The *Poisson distribution model* by Maher (1982) models the number of goals scored by each team in a match [13].
- Dixon and Coles (1997) introduced modifications to the basic Poisson model to account for low scoring games. Their approach has since been widely used in soccer prediction because of its improved accuracy [5].
- Baio and Blangiardo (2010) utilized a *Bayesian Hierarchical Model* to predict Italian Serie A match outcomes, considering team strengths and home field advantage [1].

Building on the foundations laid by deterministic and probabilistic models, the advent of advanced computational capabilities paved the way for more sophisticated prediction techniques.

2.3 Machine Learning Models

Machine learning models represent a paradigm shift, moving away from strictly formula-based predictions. These models learn from historical data, identifying patterns and relationships that can be applied to predict future match outcomes. By adapting and refining their predictions based on new data, machine learning models offer dynamic and often more accurate forecasting.

Notable Implementations:

- Joseph et al. (2006) worked with *decision trees* and *Bayesian networks* to predict soccer match outcomes [12].
- Stübinger, Mangold, and Knoll (2020) utilized machine learning for forecasting soccer match outcomes based on match and player attributes. Through a combination of different machine learning algorithms, their approach surpassed individual methodologies in performance [16].
- Rodrigues and Pinto (2022) employed various machine learning techniques including *random forests*, *support vector machines*, and *neural*

networks to predict soccer match outcomes. Their models showcased promising performance in terms of soccer betting profits [14].

2.4 Our Hybrid Model

Our model is a fusion of deterministic and machine learning approaches. It consists of two components: a dynamic rating system and logistic regression.

The dynamic rating system, inspired by the Dolores paper [2], assigns a unique rating to each team for every competition they partake in. This system has been enhanced to incorporate league ratings.

Logistic regression, a machine learning technique, uses the difference in ratings between two teams to predict the outcome probabilities of a soccer match, namely the chances of a home win, draw, or away win.

By synergizing the deterministic rating system with the machine learning-based logistic regression, our model draws upon the strengths of both methodologies. While it is not entirely novel, it extends and refines an existing approach to address the unique challenges associated with predicting international fixtures. This approach demonstrates how existing models can be adapted and refined for international prediction, representing a meaningful contribution to the field of soccer match prediction.

Chapter 3

Data Collection, Integration and Preprocessing

Data is fundamental to predictive modeling. The accuracy and reliability of a model are directly influenced by the quality and comprehensiveness of the data used. For this study, we collected data from multiple soccer leagues and competitions, ensuring a diverse and comprehensive dataset. It's worth noting that we restricted our data collection to existing and freely available datasets, without resorting to extensive data scraping.

3.1 Datasets

We utilized three different datasets in this study, each obtained from reputable sources known for their consistent and reliable soccer data. All datasets include essential parameters like match date, league, participating teams, goals scored, and match result. Additionally, the domestic cup and international datasets provide information on the domestic leagues to which each participating team belongs. This vital information enables us to train league ratings, essential for understanding and quantifying the differences in league strengths.

The primary dataset of domestic league games was used to train the team ratings and the logistic regression model. The domestic cup data helped establish internal league ratings within a nation, improving rating initialization for teams promoted or relegated to a new league for the first time. Lastly, the international game data was used to establish league ratings between leagues from different nations, necessary for predicting international games.

Domestic League Data: The primary dataset was sourced from *football-data.co.uk* [10], a well-established platform for comprehensive soccer statistics. This dataset encompasses 34,298 matches played between 2016 and 2022

across 17 different leagues. For assessing prediction accuracy, an additional dataset of 5,813 games from the 2022/23 season of these 17 leagues was used. Notably, this dataset incorporates Bet365 betting odds, which we subsequently use as a comparative benchmark for our model's predictions in Section 7.3. The included leagues are the top-tier leagues from England, Spain, Italy, Germany, Netherlands, France, Portugal, Belgium, Türkiye, Scotland, and Greece, as well as the second-tier leagues from England, Spain, Italy, Germany, France, and Scotland.

Domestic Cup Data: Data for domestic cup games was extracted from *Kaggle's Football Data from Transfermarkt* [11], a dataset derived from one of the most comprehensive soccer databases globally. Due to the constraints of freely available data, this dataset comprises a relatively smaller sample of 638 matches, offering insights into matches where teams from different domestic leagues within the same country competed. The domestic cups in this dataset come from five countries: England, Spain, Italy, Germany, and Scotland.

International Champions League Data: For UEFA Champions League matches, our dataset was sourced from *worldfootball.net* [17], a trusted platform for global football statistics. It covers 744 matches played between 2016 and 2022. To evaluate the model's prediction accuracy, a separate dataset containing all 125 games from the 2022/23 Champions League season was used.

Once the datasets were sourced, the next crucial step was data processing.

3.2 Data Processing

The preparation of our dataset involved numerous steps, from sourcing to integrating data from different origins. Proper data processing was instrumental in ensuring the reliability and accuracy of our predictive model.

Our datasets comprised several essential entries for each game. In cases where an entry was missing, we manually verified the game's details. If the missing information could be located, it was included. Otherwise, such entries were removed to maintain the integrity of our dataset.

Not all datasets uniformly represented the full-time result of matches. Some datasets detailed the number of goals scored by each team without specifying the match outcome (Home Win, Draw, Away Win). To standardize this, we implemented a script that parsed through every game entry, comparing the goals scored by the home and away teams. Based on this comparison, we determined the match outcome and added a corresponding 'Result' entry.

Ensuring consistency in team and league names across datasets was a primary concern. Variations in team naming, such as 'PSG' being referred to as 'Paris SG' in another dataset, necessitated careful manual adjustments. For league names, we adopted a standardized format: the initial three letters are an abbreviation of the country, followed by a number indicating the league tier. For instance, 'ENG1' represents the top-tier league in England, while 'ENG2' denotes the second-tier league. This format facilitated smooth data integration.

Moreover, our model demanded knowledge of the domestic league affiliations of the participating teams for every match in the domestic cup and international datasets. To pinpoint these affiliations, we cross-referenced team names with those in our domestic league dataset. This strategy was crucial in linking teams from the international dataset with their respective domestic leagues. For instance, to adjust ratings correctly after an Arsenal vs. Barcelona match, the system needed to know Arsenal's affiliation with the English Premier League and Barcelona's with the Spanish La Liga.

Chapter 4

The Model

In this section, we will explore the workings of the model in detail. The model consists of two main components.

The first component is a dynamic rating system that assigns a unique rating to each team for every competition they participate in. It is crucial to understand that a team's rating is only meaningful within the context of the specific competition for which it is assigned. After each match, the ratings of both teams are updated based on several parameters, which will be detailed in section 4.1.1.

This rating system is inspired by the one proposed in the Dolores paper [2] but is extended to include league ratings. In scenarios involving international games or domestic cup games between teams from different leagues, we update the corresponding league's ratings instead of the teams' ratings.

This approach enables us to predict international fixtures by combining a team's internal league rating with its league's rating to create a new rating that is applicable in the international context. We utilize domestic cup outcome data to assign separate league ratings within the context of their nation. This provides a more nuanced understanding of the differences in levels between leagues within the same nation, enabling us to propose an improved initialization for teams that are promoted or relegated to a new league for the first time.

Once the final ratings for both teams are determined, we compute the rating difference. This leads us to the second component of the model - the logistic regression. This part of the model takes the rating difference as input and learns to predict the outcome probabilities, which are the probabilities of a home win, draw, and away win.

4.1 The rating system

Each team has a rating that includes a home rating for when they play at home, an away rating for when they play away, and a counter for continuous over/underperformances, which counts the number of consecutive games a team has over- or underperformed. When a team consistently overperforms or underperforms for a specified number of consecutive games, its background rating is temporarily replaced with a provisional rating. This provisional rating is more responsive to the team's recent performances, allowing for a more dynamic adjustment of the team's rating.

Similarly, each domestic league has the same rating instances. For simplicity, we will explain the ratings for a team, but the league ratings work in the same way.

4.1.1 The parameters

The rating system consists of ten parameters. Below are all the parameters explained in more depth.

The original pi-rating [3], which is the foundation of the Dolores rating system, included three parameters: the team learning rate $\lambda_{\mathcal{T}}$, the diminishing function ψ and the team learning rate $\gamma_{\mathcal{T}}$. It is important to note that the team learning rates are not team-specific but are the same for all teams.

Team learning rate $\lambda_{\mathcal{T}}$ A team's predicted performance in a fixture is based on its rating prior to the game. After observing the actual result we can determine how much the team has over- or underperformed based on our expectation. The learning rate $\lambda_{\mathcal{T}}$ then determines how much a team's rating will be updated. A higher value places more weight on recent games, reflecting that recent performances are more valuable than older ones.

Diminishing function φ It can be argued that in soccer, the difference between winning by four or five goals is not as significant as the difference between drawing or winning by one goal. The diminishing function φ reflects this by diminishing the impact of each additional goal scored on the change in a team's rating. It is a function of the error between the predicted and actual goal difference of a match. The diminishing function φ is multiplied with the learning rate $\lambda_{\mathcal{T}}$ and then added to a team's previous rating to form the new team rating.

Team learning rate $\gamma_{\mathcal{T}}$ Each team has two ratings: one for home games and one for away games. However, even if a team plays at home, its away rating should also be updated, and vice versa. The learning rate $\gamma_{\mathcal{T}}$ determines how much. A larger value corresponds to a bigger influence. This learning

rate accounts for the home advantage, which is the phenomenon of teams usually performing better at home.

In the Dolores paper, three more parameters were added: the form threshold φ, the rating impact μ and the diminishing factor δ. These parameters capture a team's form factor and make the rating system more adaptable to recent performances in case of a winning or losing streak.

Form threshold φ This is the number of consecutive games a team needs to over- or underperform to trigger the provisional rating. More precisely, a team needs a streak of at least $\varphi + 1$ games for the provisional rating to be considered. In the Dolores paper, φ was set to 1, meaning that any team over- or underperforming at least 2 times in a row triggers the form factor.

Rating impact μ This parameter determines how much the provisional rating differs from the background rating. It is multiplied with a term consisting of the current over/underperformances, the form threshold φ and the diminishing factor δ and then added to or subtracted from (depending on whether a team is over- or underperforming) a team's background rating to establish its provisional rating.

The diminishing factor δ As the background ratings must eventually converge to the provisional ratings, the form impact needs to decrease with each consecutive over/underperformance. Therefore, the diminishing factor δ is introduced, which determines how much the rating impact μ decreases for each additional game in the streak.

We have added five more parameters: separate league learning rates $\lambda_{\mathcal{L}}$ and $\gamma_{\mathcal{L}}$ for league ratings, rates ρ and σ which combine a team's rating with its corresponding league rating to allow for international predictions and finally α, which is used for initializing a team's rating in a new league after promotion or relegation.

League learning rates $\lambda_{\mathcal{L}}$ and $\gamma_{\mathcal{L}}$ These are a logical consequence of introducing league ratings. While they are used in the same manner as the team learning rates, their corresponding values differ. The amount a rating needs to be updated depending on the outcome of a game might differ for team ratings as opposed to league ratings. Therefore, separate league learning rates have been introduced.

Rates ρ and σ These are needed for predicting the outcome of international games. In this scenario, we need to consider both the team's domestic league internal rating and the rating of the domestic league itself. The rates ρ and

σ determine how much the team rating and the league rating, respectively, impact the final international rating.

Rate α This rate is used to initialize a team's rating in a new league when it gets promoted or relegated. Specifically, α is used to compute the initial rating of a team in the new league based on its rating in the previous league and the ratings of the old and new leagues.

4.1.2 Calculating the Rating Difference

Predicting the outcome of a game begins with calculating the rating difference between the two teams.

First, we must determine whether one of the teams is triggering the form factor, in which case we would use the team's provisional rating instead of its background rating. To do this, we check whether the absolute value of the current count of continuous over/underperformances $\varphi_{c\tau}$ is greater than the form threshold φ, i.e., if $|\varphi_{c\tau}| > \varphi$. If so, we consider that team's provisional rating pr, which is calculated from the background rating br as:

$$pr = br + \left(\mu \times \frac{\varphi_{c\tau} - \varphi}{(\varphi_{c\tau} - \varphi)^{\delta}}\right) \text{ if } \varphi_{c\tau} > \varphi \text{ (overperformance)}$$
$$pr = br - \left(\mu \times \frac{\varphi_{c\tau} - \varphi}{(\varphi_{c\tau} - \varphi)^{\delta}}\right) \text{ if } \varphi_{c\tau} < -\varphi \text{ (underperformance)}$$

If the form threshold is not reached, we set the provisional ratings to equal their respective background ratings.

$$pr = br$$

Next, the rating difference RD is calculated as the difference between the home rating of the home team x and the away rating of the away team y.

$$RD = pr_{xH} - pr_{yA}$$

If the game we are trying to predict is an international game, we need to take an additional step before calculating the rating difference. For team τ in domestic league l, the international rating ir_{τ} can be computed from its provisional rating pr_{τ} and the league's rating lr_{l} as follows:

$$ir_{\tau} = \rho \times pr_{\tau} + \sigma \times lr_{l}$$

In this case, we would use each team's international rating to calculate the rating difference.

$$RD = ir_{xH} - ir_{yA}$$

Now, we can use the trained model to receive the outcome probabilities from the rating difference.

4.1.3 Ratings Update Theory

To comprehend the workings of the ratings update, let's consider a game where home team x plays against away team y. Prior to the game, each team τ has a home background rating $br_{\tau H}$ and an away background rating $br_{\tau A}$.

For international games, we update the domestic league ratings instead of the team ratings, and use the league learning rates $\lambda_{\mathcal{L}}$ and $\gamma_{\mathcal{L}}$ instead of the team learning rates $\lambda_{\mathcal{T}}$ and $\gamma_{\mathcal{T}}$. However, the logic remains the same.

The reason we do not update the team ratings based on international games is that team ratings are context-specific to the domestic league in which a team plays. Hence, they should only be adjusted based on the performance in games within that league. In any given league match, when one team's rating increases due to overperformance, another team's rating in the same league decreases by the same amount due to underperformance. If we were to update team ratings based on international games, we would end up increasing a team's rating in one league while decreasing another team's rating in a different league, potentially leading to inflated ratings. This would compromise the ability to predict how many goals a team would score against the average team in its league, a crucial step in updating team ratings.

On the other hand, updating the league ratings is essential for translating team ratings from different leagues to predict international games. League ratings provide the context for inter-league comparisons and enable us to assess the relative strengths of different leagues.

The ratings update process begins by observing the actual goal difference g_o, which is the difference between the goals scored by the home and away teams.

$$g_o = g_{ox} - g_{oy}$$

We then predict the goal difference of each team against the average opponent in the same league based on their background ratings. g_{px} is the predicted goal difference of home team x based on its previous home background rating $br_{xH_{t-1}}$. Similarly, g_{py} is the predicted goal difference of away team y

based on its previous away background rating $br_{yA_{t-1}}$. We use the proposed values of $b = 10$ and $c = 3$.

$$g_{px} = sgn(br_{xH_{t-1}}) \times (b^{\frac{|br_{xH_{t-1}}|}{c}} - 1)$$
$$g_{py} = sgn(br_{yA_{t-1}}) \times (b^{\frac{|br_{yA_{t-1}}|}{c}} - 1)$$

Next, we calculate the expected goal difference g_p for the match.

$$g_p = g_{px} - g_{py}$$

Now that we have both the actual and predicted goal differences, we can calculate the error e.

$$e = |g_o - g_p|$$

We input this error into the diminishing function ψ, which reduces the importance of the goal difference error. This will be needed later to update the ratings.

$$\psi(e) = c \times \log_b(1 + e)$$

In the next section, we calculate $\psi_x(e)$ and $\psi_y(e)$ from $\psi(e)$. $\psi_x(e)$ always equals $-\psi_y(e)$. The sign difference determines whether a team's background rating will increase or decrease, as you will see later. When a team overperforms, the sign is positive and its rating increases. Conversely, if a team underperforms, the sign is negative and its rating decreases.

The function $\psi_x(e)$ is defined as:

$$\psi_x(e) = \begin{cases} \psi(e), & \text{if } g_p < g_o \\ -\psi(e), & \text{otherwise} \end{cases}$$

Similarly, the function $\psi_y(e)$ is defined as:

$$\psi_y(e) = \begin{cases} \psi(e), & \text{if } g_p > g_o \\ -\psi(e), & \text{otherwise} \end{cases}$$

Finally, we can update each team's home and away background ratings.

The new home background rating of home team x (br_{xH_t}) based on its prior home background rating ($br_{xH_{t-1}}$) is calculated as follows:

$$br_{xH_t} = br_{xH_{t-1}} + \psi_x(e) \times \lambda_{\mathcal{T}}$$

The new away background rating of home team x (br_{xA_t}) based on its prior away background rating ($br_{xA_{t-1}}$) is calculated as follows:

$$br_{xA_t} = br_{xA_{t-1}} + (br_{xH_t} - br_{xH_{t-1}}) \times \gamma_{\mathcal{T}}$$

The new away background rating of away team y (br_{yA_t}) based on its prior away background rating ($br_{yA_{t-1}}$) is calculated as follows:

$$br_{yA_t} = br_{yA_{t-1}} + \psi_y(e) \times \lambda_{\mathcal{T}}$$

The new home background rating of away team y (br_{yH_t}) based on its prior home background rating ($br_{yH_{t-1}}$) is calculated as follows:

$$br_{yH_t} = br_{yH_{t-1}} + (br_{yA_t} - br_{yA_{t-1}}) \times \gamma_{\mathcal{T}}$$

4.1.4 Parameter optimization

To ensure optimal prediction accuracy, it's crucial to identify the most effective parameter values. This involves evaluating a wide array of parameter combinations and selecting the set that delivers the best performance.

The challenge then lies in defining what best performance means. For our research, we've chosen to employ the Ranked Probability Score (RPS) as our metric of success. RPS is a well-established metric in the domain of soccer predictions, providing a comprehensive measure of the accuracy of probabilistic forecasts. At its core, RPS evaluates the difference between the predicted probabilities for each potential outcome and the actual match result. A lower RPS indicates a closer match between predictions and reality, making it an ideal metric for our purpose. A detailed explanation of RPS, including its calculation will be presented in chapter 7.

In this research, we employed the Grid Search optimization technique. Grid Search systematically evaluates all potential combinations of parameters within a defined range to pinpoint the combination that offers the optimal score. Our target was to achieve the lowest possible Ranked Probability Score (RPS).

Optimization for $\lambda_{\mathcal{T}}$ and $\gamma_{\mathcal{T}}$

Our initial application of Grid Search was focused on the team learning rates, $\lambda_{\mathcal{T}}$ and $\gamma_{\mathcal{T}}$. This involved exhaustively testing each combination, striving

to identify the pair that led to the most accurate background ratings, while temporarily setting aside provisional ratings.

Referring to Figure 4.1, the optimal values determined were $\lambda_{\mathcal{T}} = 0.042$ and $\gamma_{\mathcal{T}} = 0.97$. While the $\lambda_{\mathcal{T}}$ aligns with values from the Dolores rating system and its predecessor, the pi-rating system, the $\gamma_{\mathcal{T}}$ is notably elevated. This higher $\gamma_{\mathcal{T}}$ value implies a closer link between a team's home and away performances, suggesting that the traditional strong home advantage might be less distinct than previously considered.

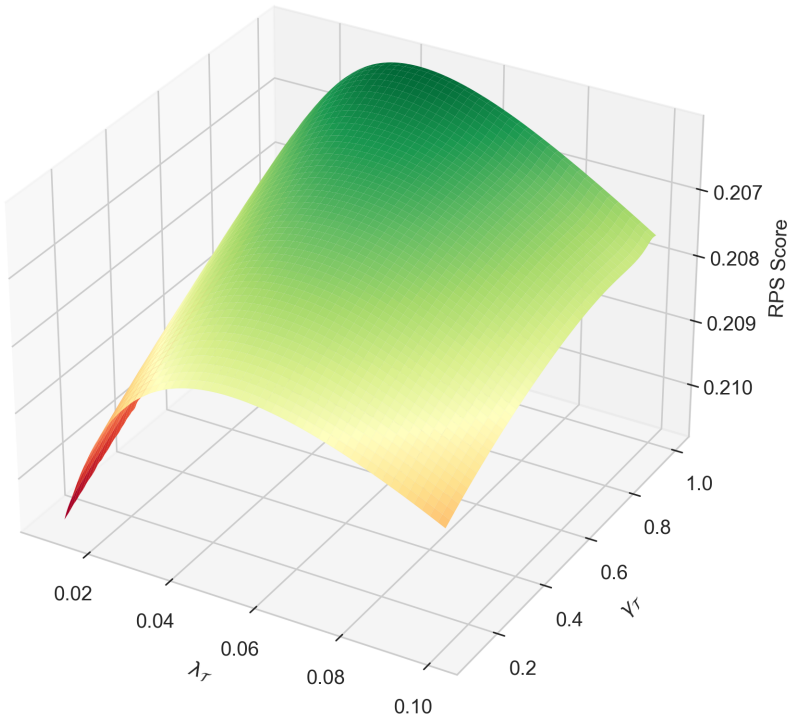

Figure 4.1: Grid Search results for team learning rates $\lambda_{\mathcal{T}}$ and $\gamma_{\mathcal{T}}$. The prediction error, quantified by the RPS, is minimized at $\lambda_{\mathcal{T}} = 0.042$ and $\gamma_{\mathcal{T}} = 0.97$. The corresponding optimal RPS value is 0.205996.

Optimization for φ, μ and δ

After determining the optimal values for the team learning rates, we proceeded to optimize the three parameters that define the provisional ratings: the form threshold φ, the rating impact μ, and the diminishing factor δ. Interestingly, the optimal values were found when $\mu = 0$, which means that the provisional ratings equaled the background ratings. This result suggests that, in our dataset, the provisional ratings do not positively impact prediction accuracy. Consequently, provisional ratings will be disregarded in subsequent analyses.

Optimization for $\lambda_{\mathcal{L}}$ and $\gamma_{\mathcal{L}}$

With the rates in the domestic league context fixed, we moved to the second stage of the analysis, where we aimed to determine the optimal league learning rates and the optimal method for combining team and league ratings to enhance the accuracy of international game predictions.

We performed another Grid Search, this time on the team learning rates $\lambda_{\mathcal{L}}$ and $\gamma_{\mathcal{L}}$. These rates function similarly to the earlier team learning rates but dictate how rapidly the league ratings are updated when teams from different domestic leagues compete. Prior to the optimization the ρ and σ values have been set to 0.75. As illustrated in Figure 4.2, the RPS for international games was minimized at $\lambda_{\mathcal{L}} = 0.13$ and $\gamma_{\mathcal{L}} = 0.96$. The found $\gamma_{\mathcal{L}}$ value is almost identical to its analogue in the team rating context, while the $\lambda_{\mathcal{L}}$ value is significantly higher. This discrepancy could be attributed to the limited international data available, meaning that the ratings need to converge more rapidly. The $\lambda_{\mathcal{L}}$ value might also be higher because a league's rating represents all its teams, and teams' strengths can vary a lot within the same league. In international games, a strong team from a league might play one match, and a much weaker team from the same league might play the next. This means the model needs to adjust league ratings more quickly to accurately reflect the teams actually playing in each international match.

Optimization for ρ and σ

In predicting international games, accurately combining a team's rating with its domestic league's rating is crucial. After experimenting with various approaches for combining these ratings, the most effective method turned out to be one of the simplest: calculating the international rating of a team as the weighted sum of the two ratings, $ir_\tau = \rho \times pr_\tau + \sigma \times lr_1$. Figure 4.3 indicates that the optimal values for this calculation were $\rho = 0.86$ and $\sigma = 0.69$.

This completes the parameter optimization process. The identified parameters yield the most accurate predictions and will be used in all subsequent

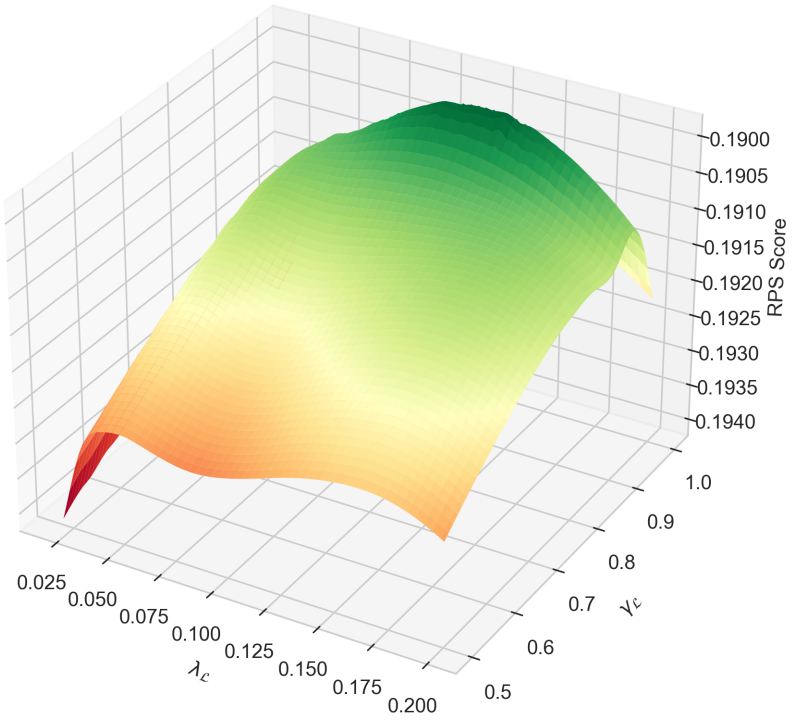

Figure 4.2: Grid Search results for league learning rates $\lambda_{\mathcal{L}}$ and $\gamma_{\mathcal{L}}$. The prediction error, quantified by the RPS, is minimized at $\lambda_{\mathcal{L}} = 0.13$ and $\gamma_{\mathcal{L}} = 0.96$. The corresponding optimal RPS value for international games is 0.189789.

analyses.

4.1.5 New Initialization approach

In the original approach, a team received a rating of 0 when it first entered a new league, regardless of its previous rating. This approach does not account for the team's performance in the previous league, which is a valuable piece of information for predicting its performance in the new league. The goal of the new approach is to use the team's rating in the old league and the ratings of the old and new leagues to predict the team's rating in the new league.

We use the same idea as with the international rating but this time in a

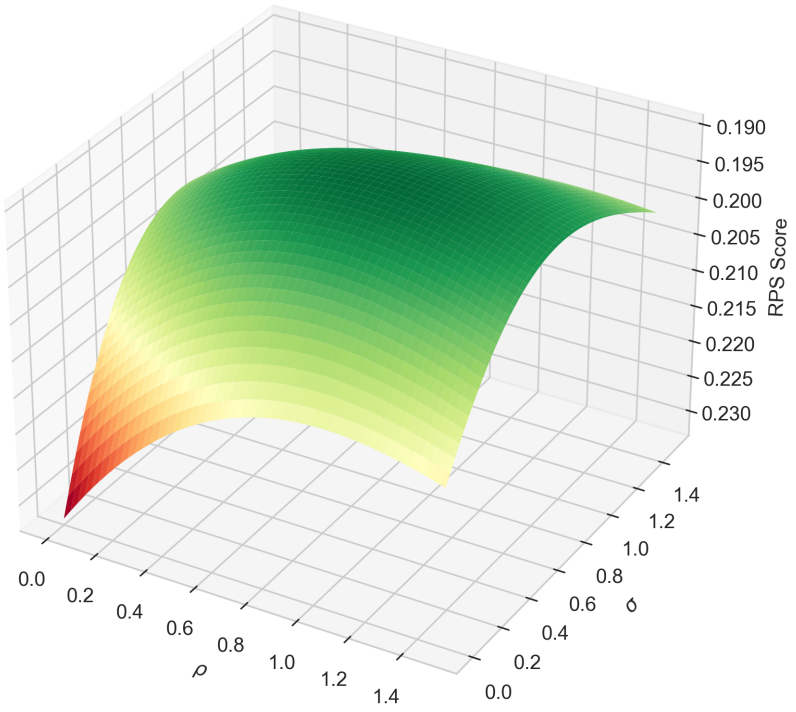

Figure 4.3: Grid Search results for rates ρ and σ. The prediction error, quantified by the RPS, is minimized at $\rho = 0.86$ and $\sigma = 0.69$. The corresponding optimal RPS value for international games is 0.189525.

different context. First, we train the different leagues' ratings from a certain nation through domestic cup games. Here, teams from different leagues but the same nation play against each other. This allows us to assign league ratings to each league through the same process as with the international ratings but this time in a nation-internal context. We do this by looking at domestic cup games and, instead of updating the team ratings after a game, we update the teams' league ratings, same as for the international league context.

Now that we have the team's rating in the old league as well as the old and new league rating, we are able to compute the team's rating in the new league by computing its national rating, that is, the rating used in the context

of games between teams from the same nation but different leagues. The concept is the same as for the international rating but note that ρ and σ do not have the same values as in the international context.

We can compute a team's national rating nr using its background rating for the old league $br_{l_{old}}$ and that league's rating $lr_{l_{old}}$.

$$nr = \rho \times br_{l_{old}} + \sigma \times lr_{l_{old}}$$

But at the same time, we could compute the national rating nr using the team's rating in the new league $br_{l_{new}}$ and that league's rating $lr_{l_{new}}$.

$$nr = \rho \times br_{l_{new}} + \sigma \times lr_{l_{new}}$$

We can now compute the team rating for the new league by combining the two equations.

$$\rho \times br_{l_{new}} + \sigma \times lr_{l_{new}} = \rho \times br_{l_{old}} + \sigma \times lr_{l_{old}}$$

which we can convert to

$$br_{l_{new}} = br_{l_{old}} + \frac{\sigma}{\rho} \times (lr_{l_{old}} - lr_{l_{new}})$$

Now we set $\frac{\sigma}{\rho}$ to α, which allows us to only have to optimize one parameter. We can optimize this parameter separately from the ρ and σ used in the context of international ratings. This gives us the final equation:

$$br_{l_{new}} = br_{l_{old}} + \alpha \times (lr_{l_{old}} - lr_{l_{new}})$$

We optimized the value of α by iterating through possible values and predicting the outcome of games. The value that minimized the RPS was chosen as the optimal value. As shown in Figure 4.4 the optimized value of α was 1.3, which reduced the RPS from 0.205996 to 0.205535. This shows that the new approach provides better prediction accuracy.

In conclusion, our new approach for initializing team ratings when they are promoted or relegated to a new league provides better prediction accuracy compared to the initial approach. However, there is still room for improvement, and future work could focus on including domestic cup data for all leagues in our dataset.

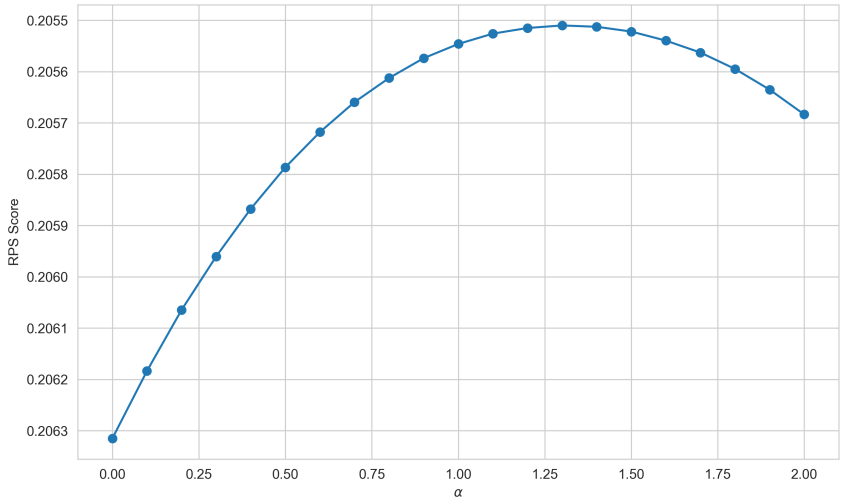

Figure 4.4: The RPS is minimized at 0.205535 for $\alpha = 1.3$.

4.2 Logistic Regression

To translate ratings into match outcome probabilities, we trained a logistic regression model using a dataset of 28'297 matches. For each match, we used the rating difference between the teams and the actual outcome (Home win, Draw, or Away win) as the input and target variables, respectively.

Logistic regression is used to predict the probability of a binary outcome based on one or more predictor variables. In our case, we used it to predict the probability of each possible match outcome (Home win, Draw, Away win) based on the rating difference between the teams. We used L2 regularization to make the model more robust, even though the risk of overfitting is quite low with only one input feature.

We also employed cross-validation to ensure the model's robustness and generalizability. The cross-validation resulted in very low variance, confirming the model's robustness and ability to generalize well to new data.

After training the model, we can use it to predict the outcome probabilities of new matches by inputting the rating difference between the teams.

Chapter 5

Demonstrating the Model: A Practical Application

This chapter tries to give the reader a better understanding of how the model works by giving a step by step walk through of a worked example. We explain how to first predict and then how to update the ratings of both domestic league and international games.

5.1 Domestic League Game

In this section, we detail the steps required to predict the outcome of a domestic league game and subsequently update the team ratings. We use the iconic match between Liverpool and Manchester United on March 5, 2023, which ended in a surprising 7:0 victory for Liverpool, as a case study.

5.1.1 Domestic Prediction

To predict the game's outcome, we must first ascertain the teams' ratings before the match. We refer to Liverpool as team x and Manchester United as team y. The prediction requires only the home rating of the home team and the away rating of the away team.

Team x home rating $br_{xH} = 0.809606$.

Team y away rating $br_{yA} = 0.565904$.

As previously mentioned, we do not use the provisional ratings at this stage, as they do not appear to enhance prediction accuracy. Therefore, we can directly compute the rating difference (RD) based on the teams' background ratings.

$$RD = br_{xH} - br_{yA} = 0.809606 - 0.565904 = 0.243703$$

We then input the value RD = 0.243703 into the logistic regression model, which yields the probabilities for a home win (H), draw (D), and away win (A).

$$P(H) = 0.4866,\ P(D) = 0.2744,\ P(A) = 0.2390$$

The model predicts a 48.66% chance of a home win, a 27.44% chance of a draw, and a 23.90% chance of an away win.

Finally, we compare our predictions with the actual outcome to measure accuracy. Our predictions indicate Liverpool as the favorites to win, but a draw or an away win also seem plausible. As previously mentioned, Liverpool dominated the game, resulting in a 7:0 victory which nobody saw coming. Our model does not attempt to predict the score, only the winner. Consequently, the resulting RPS score is 0.1603, which is relatively good.

5.1.2 Team Ratings Update

After a game has concluded, it is essential to update the teams' ratings based on the match's outcome. The home and away rating of a team τ before the game are denoted as $br_{\tau H_{t-1}}$ and $br_{\tau A_{t-1}}$, respectively. The ratings before the game were:

- Team x prior home rating $br_{xH_{t-1}} = 0.809606$.
- Team x prior away rating $br_{xA_{t-1}} = 0.728574$.
- Team y prior home rating $br_{yH_{t-1}} = 0.614804$.
- Team y prior away rating $br_{yA_{t-1}} = 0.565904$.

Our objective is to compute the posterior home and away ratings for each team, $br_{\tau H_t}$ and $br_{\tau A_t}$, respectively.

The first step is to compute the observed goal difference, which is the difference in goals scored by the home and away teams.

$$g_o = g_{ox} - g_{oy} = 7 - 0 = 7$$

Next, we predict the goal difference of each team against the average opponent in the same league, based on their background ratings. The expected goal difference against the average opponent in the league for home team x, g_{px}, and away team y, g_{py}, are calculated as follows:

$$g_{px} = sgn(br_{xH_{t-1}}) \times (b^{\frac{|br_{xH_{t-1}}|}{c}} - 1)$$
$$g_{py} = sgn(br_{yA_{t-1}}) \times (b^{\frac{|br_{yA_{t-1}}|}{c}} - 1)$$

Next, we calculate the expected goal difference g_p for the match.

$$g_{px} = sgn(br_{xH_{t-1}}) \times (10^{\frac{|br_{xH_{t-1}}|}{3}} - 1) = 1 \times (10^{\frac{0.809606}{3}} - 1) = 0.861525$$
$$g_{py} = sgn(br_{yA_{t-1}}) \times (10^{\frac{|br_{yA_{t-1}}|}{3}} - 1) = 1 \times (10^{\frac{0.565904}{3}} - 1) = 0.543955$$

The expected goal difference, g_p, is then the difference between g_{px} and g_{py}.

$$g_p = g_{px} - g_{py} = 0.861525 - 0.543955 = 0.317570$$

To compute the prediction error, e, we take the absolute difference between the observed and predicted goal differences.

$$e = |g_o - g_p| = |7 - 0.307679| = 6.682430$$

We then input the error, e, into the diminishing function, ψ, which reduces the impact of the goal difference error.

$$\psi(e) = 3 \times \log_{10}(1 + e) = 3 \times \log_{10}(1 + 6.682430) = 2.656496$$

Next, we compute the adjusted errors, $\psi_x(e)$ and $\psi_y(e)$, from $\psi(e)$. Since $g_p < g_o$, $\psi_x(e)$ and $\psi_y(e)$ are computed as follows:

$$\psi_x(e) = \psi(e) = 2.656496$$
$$\psi_y(e) = -\psi(e) = -2.656496$$

The sign of the adjusted errors determines whether a team's rating will increase or decrease. As Liverpool overperformed, their rating will increase, while Manchester United's rating will decrease due to their underperformance.

Finally, we update each team's home and away background ratings using their prior ratings, the adjusted errors, and the optimal team learning rates, $\lambda_{\mathcal{T}}$ and $\gamma_{\mathcal{T}}$, obtained from parameter optimization as described in section 4.1.4.

The new home background rating of home team x (br_{xH_t}) is calculated as:

$$\begin{aligned} br_{xH_t} &= br_{xH_{t-1}} + \psi_x(e) \times \lambda_{\mathcal{T}} \\ &= 0.809606 + 2.656496 \times 0.042 = 0.921179 \end{aligned}$$

The new away background rating of home team x (br_{xA_t}) is calculated as:

$$\begin{aligned} br_{xA_t} &= br_{xA_{t-1}} + (br_{xH_t} - br_{xH_{t-1}}) \times \gamma_{\mathcal{T}} \\ &= 0.728574 + (0.921179 - 0.809606) \times 0.97 = 0.836799 \end{aligned}$$

The new away background rating of away team y (br_{yA_t}) is calculated as:

$$\begin{aligned} br_{yA_t} &= br_{yA_{t-1}} + \psi_y(e) \times \lambda_{\mathcal{T}} \\ &= 0.565904 - 2.656496 \times 0.042 = 0.454331 \end{aligned}$$

The new home background rating of away team y (br_{yH_t}) is calculated as:

$$\begin{aligned} br_{yH_t} &= br_{yH_{t-1}} + (br_{yA_t} - br_{yA_{t-1}}) \times \gamma_{\mathcal{T}} \\ &= 0.614804 + (0.454331 - 0.565904) \times 0.97 = 0.506579 \end{aligned}$$

These new ratings can now be used to predict future games. It is important to note that the ratings adjusted significantly in this example due to the considerable overperformance of Liverpool and underperformance of Manchester United.

5.2 International Game

In this section, we will analyze one of the most anticipated matches of the 2022/23 season - the UEFA Champions League final between Manchester City and Inter Milan.

5.2.1 International Prediction

To predict the outcome of the game, we first need to look at the ratings prior to the match. Manchester City is referred to as team x and Inter Milan as team y. To predict the outcome, we only need the home rating of the home team and the away rating of the away team.

Team x home rating $br_{xH} = 1.260805$.

Team y away rating $br_{yA} = 0.810623$.

Note that in this specific case of the Champions League final, no team is technically at home or away. However, as the home and away ratings differ only slightly, it does not greatly influence the prediction. In the future, the system could be expanded to consider the average between the two ratings for each team when predicting a final.

As we are trying to predict an international fixture, we also need the respective team's domestic league ratings. We refer to Manchester City's league,

the Premier League, as league X, and to Inter Milan's league, the Serie A, as league Y. Prior to the game, the ratings are:

League X home rating $lr_{XH} = 1.659532$.

League Y away rating $lr_{YA} = 1.668503$.

The next step is to compute each team's international rating based on their team rating and their team's league rating. For team τ of league l, the international rating is computed as follows:

$$ir_\tau = \rho \times pr_\tau + \sigma \times lr_l$$

The values $\rho = 0.86$ and $\sigma = 0.69$ are obtained from the parameter optimization in section 4.1.4. We can now plug in all the values to compute each team's international rating.

$$\begin{aligned} ir_{xH} = \rho \times br_{xH} + \sigma \times lr_{XH} &= 0.86 \times 1.260805 + 0.69 \times 1.659532 = 2.229370 \\ ir_{yA} = \rho \times br_{yA} + \sigma \times lr_{YA} &= 0.86 \times 0.810623 + 0.69 \times 1.668503 = 1.848403 \end{aligned}$$

We proceed by calculating the rating difference (RD) for an international game.

$$RD = ir_{xH} - ir_{yA} = 2.229370 - 1.848403 = 0.380967$$

We can now input the value RD $= 0.380967$ into the logistic regression model, and we receive the probabilities for home win (H), draw (D), and away win (A). These probabilities refer to the result after 90 minutes, otherwise there could not be a draw in a final.

$$P(H) = 0.5215,\ P(D) = 0.2669,\ P(A) = 0.2116$$

The model predicts a 52.15% chance of a home win, a 26.69% chance of a draw, and a 21.16% chance of an away win.

Finally, we can compare our predictions to the actual outcome and measure the accuracy. Our predictions suggest quite confidently that Manchester City will win the game. As you might know, Manchester City indeed won the Champions League final. As a consequence, the resulting RPS score is 0.1368, which is low. This suggests that our model's prediction was quite accurate and is consistent with the actual outcome.

5.2.2 League Ratings Update

After an international game, instead of updating the team ratings, we update the ratings of the teams' domestic leagues. The process is exactly the same as for updating team ratings, but we use the league ratings instead of the team ratings and the league learning rates, $\lambda_{\mathcal{L}} = 0.13$ and $\gamma_{\mathcal{L}} = 0.96$, instead of the team learning rates, $\lambda_{\mathcal{T}} = 0.042$ and $\gamma_{\mathcal{T}} = 0.97$. Therefore, refer to section 5.1.2 to understand how ratings are updated after a game.

Chapter 6

Implementation

6.1 System Overview

The system implemented is engineered to forecast soccer match outcomes and revise the ratings of teams and leagues based on the actual match results. The model considers both the existing ratings of the participating teams and the league in which they are competing to forecast a match's outcome. Subsequently, after the match concludes, the system updates the teams' and leagues' ratings based on the actual result.

6.2 Environment and Libraries

The system is developed using Python 3.9 and incorporates several external libraries to streamline the implementation:

- `numpy`: Employed for numerical computations.
- `pandas`: Utilized for data manipulation and analysis.
- `math`: Employed for mathematical operations.
- `scipy`: Utilized for parameter optimization.
- `sklearn`: Employed for the logistic regression.
- `matplotlib`: Utilized for plotting and visualization.

Each library was selected for its specialized functionality utilized within the system.

6.3 Data

The system leverages historical match data, including the participating teams, the league in which they are competing, and the actual match outcome. The

data is cleaned and preprocessed to ensure it is appropriately formatted for the model.

The data is stored in multiple CSV files. One for domestic league games, another for domestic cup games and finally one for international games.

For domestic league games, each game has the entries Div, Date, HomeTeam, AwayTeam, FTHG (number of goals scored by the home team at full time), FTAG (number of goals scored by the away team at full time), and FTR (the full-time result, which is H for home win, D for draw, and A for away win).

The domestic cup and international game data sheets each have two additional entries, HomeDiv (the domestic league in which the home team competes) and AwayDiv (the domestic league in which the away team competes), which are necessary to update the teams' corresponding league ratings.

6.4 Key Functions

The system is structured into several key functions, each responsible for a specific part of the process:

- `initialize_ratings`: This function initializes the ratings for all teams and leagues at the start of the process. The ratings of each team and league are initially set to 0.
- `update_ratings`: This function updates the ratings of the teams based on the actual match outcome. It is first used on a large number of games to generate meaningful ratings. Then, after each predicted match, the ratings of the two involved teams are updated using the actual outcome.
- `update_league_ratings`: Used similarly to the `update_ratings` function, but in the context of league ratings.
- `train_model`: This function trains the model based on historical data of rating differences and match outcomes.
- `predict_outcome`: This function predicts the outcome of domestic league matches based on the current ratings of the participating teams. It uses the `calculate_rating_difference` function and then inputs the rating difference as well as the trained model in a function called `calculate_probabilities` which then calculates and returns the outcome probabilities. Finally, `predict_outcome` compares the predicted outcomes to the actual outcome and returns the corresponding RPS.
- `predict_outcome_int`: This function predicts the outcome of international matches. Here, an extra step is needed to calculate each team's international rating before computing the rating difference.

- `calculate_rating_difference`: This function takes the ratings of two teams in the context and calculates their rating difference. It uses the `calculate_provisional_rating` as a subroutine.
- `calculate_provisional_rating`: This function checks whether a team is on a streak that would trigger the provisional rating and, in that case, calculates and returns that rating.
- `calculate_rps`: This function calculates the RPS for a match based on the predicted probabilities and the actual outcome.

Each function is invoked in the main function, which orchestrates the entire process.

6.5 Model Implementation

The system implements a logistic regression model using the sklearn library with an L2 penalty and the saga solver.

6.6 Parameter Optimization

The system includes functions for optimizing the model parameters using a Grid Search approach. This includes functions to optimize the team learning rates, league learning rates, and the rates that combine them.

6.7 Results and Visualization

The system includes functions for generating and visualizing the prediction process and results. This includes functions for plotting the ratings of teams and leagues over time, the RPS distribution, as well as the plots of the parameter optimizations.

Chapter 7

Evaluation and Discussion

This chapter evaluates the model's performance in terms of prediction accuracy using the Ranked Probability Score (RPS) and discusses the implications of the findings. The RPS is particularly well-suited to events with three or more possible outcomes, such as soccer matches (home win, draw, away win). Additionally, in section 7.3, we will explore a comparison of our model's predictions with normalized betting odds from Bet365, offering an external benchmark to evaluate our model's effectiveness.

7.1 Understanding the Ranked Probability Score

The RPS quantifies the accuracy of probabilistic predictions by accounting for the difference between predicted probabilities and actual outcomes across all possible results. It assigns penalties for inaccuracies in the prediction, with larger penalties for predictions further from the actual outcome. A perfect prediction will have an RPS of 0, while the maximum RPS is 1. Generally, a lower RPS indicates a more accurate prediction.

The formula for RPS is:

$$RPS = \frac{1}{r-1} \sum_{j=1}^{r-1} \left(\sum_{i=1}^{j} p_i - o_i \right)^2$$

Where:

- r is the total number of possible outcomes (3 in our case: home win, draw, away win),
- p_i represents the predicted probability of the outcome at position i,
- o_i denotes the actual outcome at position i.

For example, consider a scenario where our model predicts the probabilities for a home win, draw, and away win as 0.5, 0.3, and 0.2 respectively. If the actual result is a draw, the RPS is calculated as:

$$\frac{1}{2}\sum_{j=1}^{2}\left(\sum_{i=1}^{j} p_i - o_i\right)^2 = \frac{1}{2}\left((0.5-0)^2 + (0.8-1)^2\right) = \frac{1}{2}(0.25+0.04) = 0.145$$

In this case, the RPS is 0.145. Lower values of RPS indicate better model performance.

7.2 Model Evaluation based on RPS

7.2.1 Comparison of RPS Distributions

Figure 7.1 illustrates the RPS distribution of predicted domestic league games compared to predicted international games. The two distributions are similar, except that the international prediction has a relatively larger number of games with very low RPS. This suggests that the model is quite accurate in both cases, but especially so for international games with a significant disparity in team skill levels.

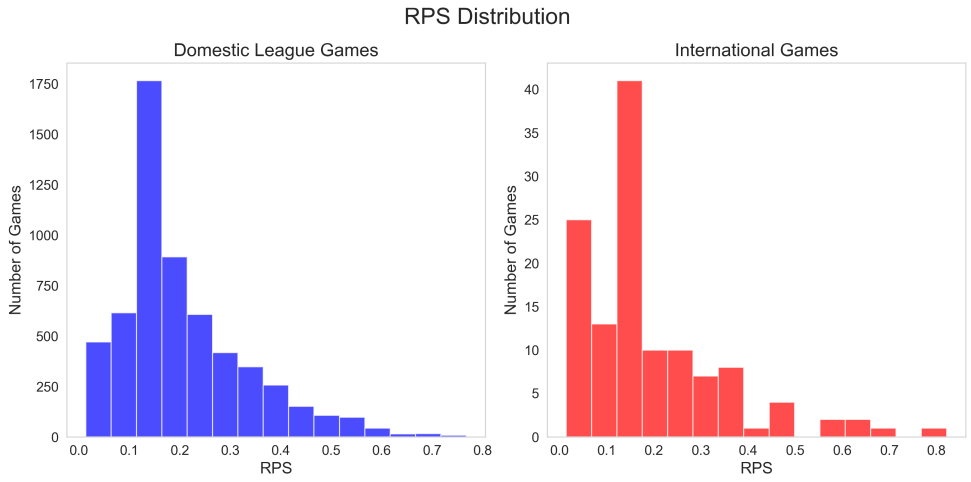

Figure 7.1: RPS distribution of predicted domestic league games and predicted international games.

7.2.2 Impact of Improved Initialization

The implementation of the improved initialization approach for newly promoted or relegated teams resulted in a decrease in the average RPS across all

17 domestic leagues, from 0.205996 to 0.205535. This improvement indicates the effectiveness of the new approach. However, the impact could be even more significant with the inclusion of additional data for second-tier leagues and domestic cups. Currently, our dataset only includes the second-tier data of six leagues and the domestic cup competitions of five nations. By incorporating a more extensive and comprehensive dataset, we can establish more accurate nation-internal league ratings, therefore eliminating the need for 0 initializations and potentially further reducing the RPS.

7.2.3 Performance Across Different Leagues

Figure 7.2 displays the average RPS for each predicted league, alongside the overall average RPS, the average RPS for top leagues, and the average RPS for second-tier leagues. It is evident from the figure that the model exhibits superior performance in predicting games from the top leagues of each country compared to the second-tier leagues. This discrepancy in performance can be attributed to the similarity in ratings among teams in the lower leagues, which consequently leads to a higher degree of uncertainty in the predictions.

7.2.4 Comparison with Dolores Model

Comparing our model with the Dolores model [2] is not straightforward. While we used a simple logistic regression approach, Dolores opted for a Bayesian Network. The Dolores model recorded an average RPS of 0.208256 across 26 leagues in a machine learning competition. However, since we used different datasets for training and testing, a direct comparison can be misleading.

In terms of mechanics, our rating system and the one from the Dolores model are very similar for domestic league predictions. The main differences are in the chosen parameters. For instance, we didn't use provisional ratings in our model because they didn't improve accuracy with our data, while Dolores might have found them useful with their dataset. We also introduced a new way to initialize ratings for teams moving between leagues, which Dolores didn't use.

While we could compare the two models by looking at parameter choices, this might not give meaningful insights. If we applied Dolores' parameters to our dataset, we'd probably get worse results, as those parameters were fine-tuned for a different data set.

We did notice a clear improvement in our model's performance when we introduced the new way to initialize ratings for teams transitioning leagues as we discussed in section 7.2.2. This shows that our model has some advantages

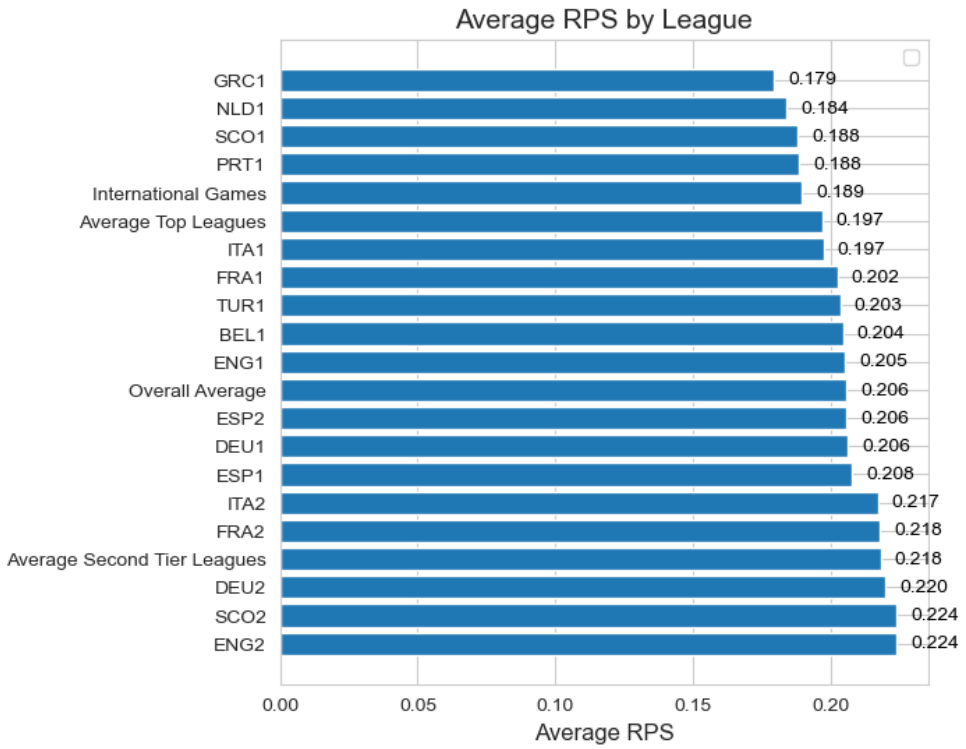

Figure 7.2: Average RPS per league, sorted from best to worst, including the overall average RPS across all leagues, the average RPS for top leagues, the average RPS for second-tier leagues, and the RPS for international games.

over the Dolores model and that our new method could be useful for other prediction models in the future.

Lastly, it's essential to emphasize that surpassing the Dolores score is not the primary objective of this study. One goal is to demonstrate that a system like Dolores can be enhanced with a new initialization approach for teams that get promoted or relegated. This can be proven without directly comparing our model with Dolores. If our model's prediction accuracy increases with the introduction of this new initialization method, it's reasonable to assume a similar improvement would occur in the Dolores model and other related rating systems.

Furthermore, this study aims to show that a system like Dolores can be extended with league ratings to predict international games. In this regard, a comparison with Dolores is inherently impossible since Dolores doesn't have the capability to predict international matches.

7.2.5 International Prediction through League Ratings

The introduction of international prediction through league ratings was successful, resulting in an RPS of 0.189367 for the 2022/23 Champions League season after implementing the new initialization approach. This score is even higher than the average RPS for the top leagues as we can see in Figure 7.2, although it might also be influenced by the large difference in skill in certain international match-ups.

7.3 Comparison with Normalized Betting Odds

Betting odds provide an insightful benchmark for evaluating predictive models. Popular platforms such as Bet365 derive their odds from various factors, providing a reliable perspective on anticipated match outcomes. Our dataset, which we obtained from football-data.co.uk [10], contained Bet365 odds. In our analysis, we compared our model's forecasts directly against these odds for every domestic league match in our prediction set that featured them, encompassing a total of 5774 games.

Betting odds are a reflection of the likelihood of a particular outcome. However, they are not directly presented as probabilities. To convert these odds into implied probabilities, one can use the formula:

$$P_i = \frac{1}{o_i}$$

Where P_i is the implied probability of outcome i and o_i is the corresponding odds offered by the bookmaker.

However, an inherent feature of betting odds is the inclusion of a margin by bookmakers, known as the overround. This ensures that, on average, they profit regardless of the outcome. As a result of this margin, the sum of the implied probabilities for all possible outcomes frequently exceeds 1.

To adjust for this, the implied probabilities are normalized to ensure their sum is exactly 1. This is achieved with the formula:

$$P_{i,\text{normalized}} = \frac{P_i}{\sum_j P_j}$$

Where $P_{i,\text{normalized}}$ is the normalized probability for outcome i and the denominator represents the sum of the implied probabilities for all outcomes.

The Mean Absolute Error (MAE) is a widely-utilized metric for evaluating prediction accuracy. It calculates the average absolute difference between predicted and actual values:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Where y_i is the probability from normalized betting odds, and $\hat{y}_i$ is our model's prediction. For the analyzed games, the MAE values between our predictions and the normalized odds were:

- Home Win: 0.0500
- Draw: 0.0199
- Away Win: 0.0454

Each value denotes the average magnitude of the difference between our model's predicted probabilities and the implied probabilities derived from Bet365's odds, for the three possible outcomes of a football match.

For instance, an MAE of 0.0500 for 'Home Win' means that, on average, our model's predicted probability for a home team winning deviates by 5.00% from the betting odds implied probability.

Given the relatively low values of the MAE across all outcomes, it is evident that our model's predictions are in close agreement with the market expectations as reflected in the Bet365 odds. Such a close alignment suggests that our model captures the same underlying dynamics and factors that the betting market, one of the most informed aggregators of such information, takes into account.

As depicted in Figure 7.3, each data point represents a specific football match, plotted by comparing the normalized betting odds' implied probability on the x-axis with the predicted probability from our model on the y-axis. The central line signifies perfect agreement, where the model's predictions and the betting odds would match exactly.

The even distribution of data points around the central line underscores the close alignment between our model's predictions and the market consensus as represented by Bet365. Moreover, the presence of a continuous stretch of data points along this line, spanning from lower to higher probabilities, indicates that our model's predictions align consistently with the betting odds across the entire probability spectrum. This consistent alignment, irrespective of whether the match is perceived as closely contested or having a clear favorite, showcases the model's capability to capture the nuances of diverse football match outcomes.

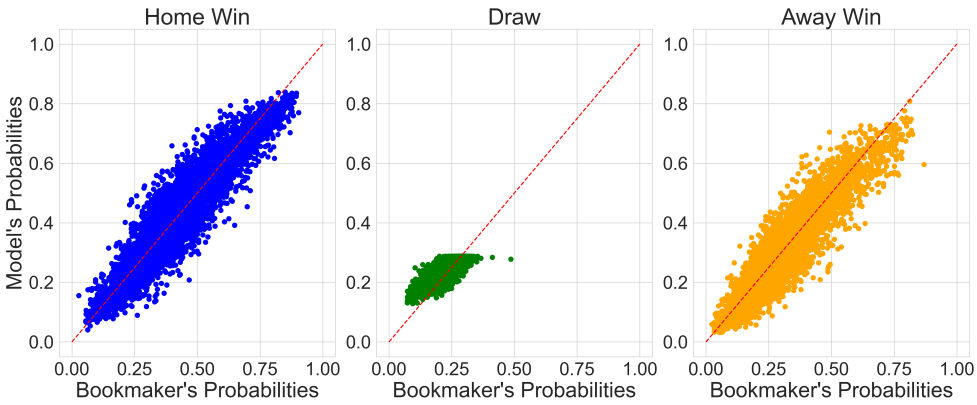

Figure 7.3: Comparison between our model's predicted probabilities and the normalized betting odds from Bet365 for home wins, draws, and away wins.

7.4 Rating Development

In this section, we visualize and discuss the evolution of team and league ratings over time, from the 2016/17 season up to and including the 2021/2022 season.

Figure 7.4 shows the team ratings over time for the 'Big Six' teams of England: Arsenal, Manchester City, Manchester United, Liverpool, Chelsea, and Tottenham. The ratings are calculated as the average of the home and away rating of the teams after each game they played. The x-axis represents the 'Number of Games Played by Team', which means each point on the x-axis corresponds to a game played by a specific team.

Figure 7.5 shows the league ratings over time that are used for predicting international games for the 11 top leagues that we also have in our domestic league dataset. The ratings are calculated as the average of the home and away rating of the leagues after each international game involving teams from those leagues. The x-axis represents the 'Total International Games Played', which means each point on the x-axis corresponds to an international game played, not specific to any league.

It is important to clarify the reason behind using different x-axes for the two figures. In the case of team ratings in domestic leagues, all teams play the same amount of games, so it is straightforward to use the 'Number of Games Played by Team' as the x-axis. However, in the case of international games, some leagues might be represented much more often than others. For example, teams from the English Premier League might be involved in more international games than teams from the Greek Super League. Therefore, using the 'Total International Games Played' as the x-axis for league ratings is more appropriate as it accounts for the varying representation of different

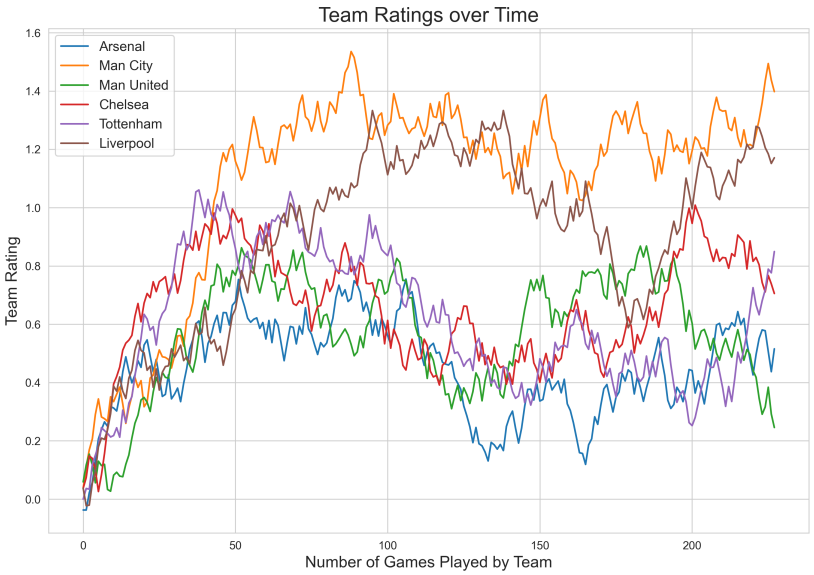

Figure 7.4: Team ratings over time for the 'Big Six' English teams. The ratings are averaged between the home and away ratings of each team after each game they played from the 2016/17 season up to and including the 2021/22 season.

leagues in international games.

Both the team and league ratings start at 0 for game day 1. The team ratings prove to be much more stable after about 50 games played per team. This is due to the much higher learning rates for league ratings, which in turn comes down to us having less data for international games; therefore, they need to adjust faster. But maybe even more importantly, a league rating represents all the teams in that corresponding league and the difference in skill by the teams represented can be quite significant. For example, if Real Madrid is the only Spanish team that makes it to the knockout stage, then the Spanish league is only represented by Real Madrid until the next tournament. Therefore, it will have quite high expectations for Spanish teams. But if then in the next season's tournament another weaker Spanish team plays in the tournament, it will almost definitely underperform drastically, and therefore the rating adapts very quickly.

A possible improvement in the future would be a larger dataset of international games, which could also include different competitions, not just the Champions League, such as the Europa League or the Conference League. This could help to create a more stable and accurate representation of the league ratings over time.

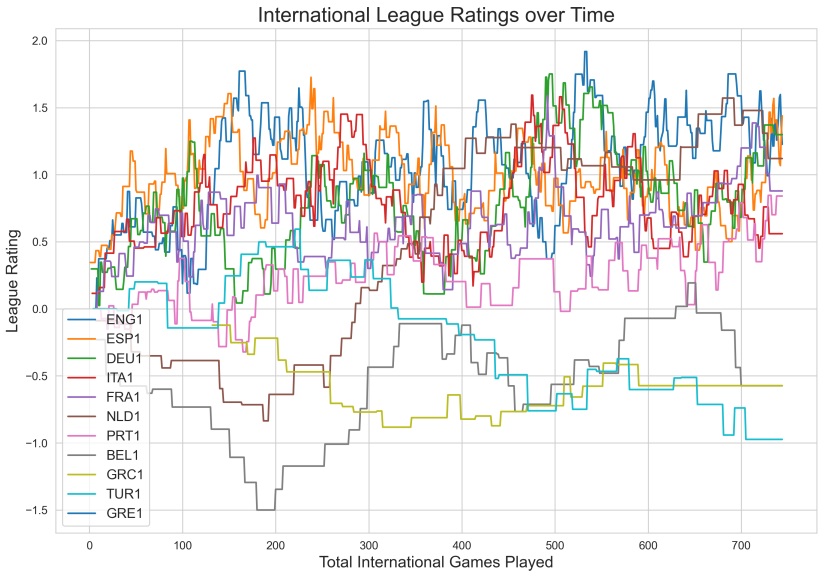

Figure 7.5: League ratings over time used for predicting international games for 11 top tier leagues. The ratings are the average of the home and away ratings of the leagues after each international game involving teams from these leagues from the 2016/17 season up to and including the 2021/22 season.

7.5 Summary of Findings

The model developed in this work demonstrates promising prediction accuracy, as evidenced by the RPS, and performs on par with the Dolores model, despite utilizing a simpler logistic regression approach and excluding provisional ratings.

The implementation of an improved initialization approach successfully reduced the average RPS across all leagues, indicating its effectiveness. Moreover, the model effectively handled the introduction of international prediction through league ratings, showcasing particularly strong performance in predicting the 2022/23 Champions League season.

A comparative evaluation with normalized betting odds from Bet365 further validated the model's predictions. This benchmarking revealed that our model's predictions are closely aligned with market expectations. The mean absolute errors between our predictions and normalized probabilities for home wins, draws, and away wins were notably low. The visual examination of the comparison further highlighted that across a spectrum of probability values, from low to high, our model's predictions consistently mirrored the market consensus, emphasizing its reliability in diverse scenarios.

Nonetheless, there are several opportunities for further improvement. Incorporating a more comprehensive dataset could facilitate the establishment of more accurate nation-internal league ratings, thereby eliminating the need for 0 initializations. Additionally, expanding the dataset to include more international games could stabilize the league ratings, potentially leading to enhanced accuracy in international predictions.

Chapter 8

Conclusion

This thesis presents an addition to the existing field of soccer match outcome prediction by developing a refined approach that builds upon and extends the rating system used in the Dolores model. While a large portion of the work was dedicated to reimplementing the existing rating system, new improvements were also incorporated. The proposed model combines a dynamic rating system to compute the rating difference between two teams and a logistic regression model to estimate the outcome probabilities based on the rating difference. This rating system leverages both team and league ratings to enhance prediction accuracy and includes nation internal league ratings to improve domestic predictions and nation overarching league ratings to enable international predictions.

The key contributions and findings of this thesis are as follows:

Improved Initialization Approach: A refined approach for initializing the ratings of newly promoted or relegated teams was proposed and implemented. This approach utilizes nation-internal league ratings to compare the strengths of leagues from the same nation. This allows us to translate the strength of a team compared to the opponents in one league to those of another league. The implementation of this approach resulted in a decrease in the average RPS across all 17 domestic leagues, indicating its effectiveness in improving prediction accuracy.

Successful International Prediction: The model successfully predicted the outcomes of the 2022/23 Champions League season. This success was achieved by incorporating nation overarching league ratings to allow for international prediction. The model achieved a higher RPS than the average for the top leagues, indicating its potential for predicting international matches with a high degree of accuracy.

The comparison with the Dolores model demonstrated that, although we opted for logistic regression instead of a Bayesian network and omitted provisional ratings, the proposed model achieved comparable prediction accuracy.

Future work could involve several directions. First, incorporating more data from lower-tier leagues could further improve the higher-tier leagues' accuracy by improving the predictions on promoted teams. It could also include training and predicting different international competitions such as the Europa League and the Conference League. Using the model to simulate entire league outcomes is another avenue worth exploring. Additionally, analyzing profitability by running a betting simulation with the model's predictions could offer valuable findings.

In summary, this thesis developed an approach aimed at improving soccer match prediction, and the results are promising. While there are areas for further improvement and exploration, the approach provides a solid foundation for future work in this area. Ultimately, this research contributes to the ongoing efforts to develop more accurate and reliable models for predicting soccer match outcomes, which has a wide range of applications from sports betting to team management and media analysis.

Bibliography

[1] Gianluca Baio and Marta Blangiardo. Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, 37:253–264, 02 2010. doi:10.1080/02664760802684177.

[2] Anthony Constantinou. Dolores: a model that predicts football match outcomes from all over the world. *Machine Learning*, 108, 01 2019. doi:10.1007/s10994-018-5703-7.

[3] Anthony Constantinou and Norman Fenton. Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal of Quantitative Analysis in Sports*, 9:37–50, 01 2013. doi:10.1515/jqas-2012-0036.

[4] Deloitte. Annual review of football finance 2023, 2023. Accessed: August 30, 2023. URL: https://www2.deloitte.com/uk/en/pages/sports-business-group/articles/annual-review-of-football-finance-europe.html.

[5] 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, 01 2002. doi:10.1111/1467-9876.00065.

[6] A.E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Pub., 1978. URL: https://books.google.ch/books?id=8pMnAQAAMAAJ.

[7] Edward S. Epstein. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology (1962-1982)*, 8(6):985–987, 1969. URL: http://www.jstor.org/stable/26174707.

[8] FIFA. Fifa big count 2006: 270 million people active in football, 2006. Accessed: September 4, 2023. URL: https://digitalhub.fifa.com/m/55621f9fdc8ea7b4/original/mzid0qmguixkcmruvema-pdf.pdf.

[9] FIFA. The fifa/coca-cola world ranking - ranking procedure, 2021. Accessed: September 4, 2023. URL: https://www.fifa.com/fifa-world-ranking/procedure/men.

[10] Football-Data.co.uk. Odds & results: Main leagues, 2023. Accessed: September 4, 2023. URL: https://www.football-data.co.uk/.

[11] Scraped from Transfermarkt and uploaded by davidcariboo. Football data from transfermarkt, 2023. Data originally sourced from Transfermarkt website and made available on Kaggle by davidcariboo. Accessed: September 4, 2023. URL: https://www.kaggle.com/datasets/davidcariboo/player-scores.

[12] Adrian Joseph, Norman Fenton, and Martin Neil. Predicting football results using bayesian nets and other machine learning techniques. *Knowl.-Based Syst.*, 19:544–553, 11 2006. doi:10.1016/j.knosys.2006.04.011.

[13] M. J. Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, September 1982. URL: https://ideas.repec.org/a/bla/stanee/v36y1982i3p109-118.html, doi:10.1111/j.1467-9574.1982.

[14] Fátima Rodrigues and Ângelo Pinto. Prediction of football match results with machine learning. *Procedia Computer Science*, 204:463–470, 2022. International Conference on Industry Sciences and Computer Science Innovation. URL: https://www.sciencedirect.com/science/article/pii/S1877050922007955, doi:https://doi.org/10.1016/j.procs.2022.08.057.

[15] Jack Rollin, Richard C. Giulianotti, Peter Christopher Alegi, Eric Weil, and Bernard Joy. football, 2023. Accessed: August 30, 2023. URL: https://www.britannica.com/sports/football-soccer.

[16] Johannes Stübinger, Benedikt Mangold, and Julian Knoll. Machine learning in football betting: Prediction of match results based on player characteristics. *Applied Sciences*, 10(1), 2020. URL: https://www.mdpi.com/2076-3417/10/1/46, doi:10.3390/app10010046.

[17] WorldFootball.net. Uefa champions league results, 2023. Accessed: September 4, 2023. URL: https://www.worldfootball.net/competition/champions-league/.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

FROM RATINGS TO RESULTS: ADVANCED PREDICTIVE MODELING FOR SOCCER OUTCOMES

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
VOGT	GILLES ISAIAH

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
ZURICH, 05.09.2023	G. Vogt

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.