# Predicting the Final League Tables of Domestic Football Leagues

Jan Van Haaren and Jesse Davis

KU Leuven, Department of Computer Science
Celestijnenlaan 200A, 3001 Leuven, Belgium
{jan.vanhaaren,jesse.davis}@cs.kuleuven.be

**Abstract**

In this paper, we investigate how accurately the final league tables of domestic football leagues can be predicted, both before the start of the season and during the course of the season. To this end, we perform an extensive empirical evaluation that compares two flavors of the well-established Elo-ratings and the recently introduced pi-ratings. We validate the different approaches using a large volume of historical match results from several European football leagues. We assess how well each ranking system performs on this task, investigate what is the most natural metric to measure the quality of a predicted final league table, and the minimum number of matches that needs to be played in order to yield useful predictions. We find that the proportion of correctly predicted relative positions is a natural metric to assess the quality of the predicted final league tables and that the traditional Elo-rating system performs well in most settings we considered.

## 1 Introduction

The prediction of football matches has received significant attention over the past few decades. Predicting the outcomes of individual football matches is a very challenging task due to the low number of goals scored [1]. Despite this fact, the ever-growing interest in football betting has this an active area of research. Initially, the focus was on purely statistical models (e.g., [3, 6, 7, 8, 9]) but in recent years it has somewhat shifted to predictive ranking systems (e.g., [2, 5]). In contrast to the large body of work in the area of predicting individual match outcomes, the related task of predicting final league tables has remained almost unexplored to date. However, the latter task is of much more interest to managers and directors who want to develop a successful long-term vision for their clubs.

In this paper, we investigate how accurately the final league tables of domestic football leagues can be predicted, both before the start of the season and during the course of the season. We focus our study on different flavors of two popular predictive ranking systems, namely the recently introduced pi-ratings [2] and the well-established Elo-ratings [5]. We validate the different systems on a large volume of historical match results covering over twenty seasons of the most important European football leagues. Besides assessing how well each predictive system performs on this task, we also investigate what is the most natural metric to measure the quality of a predicted final league table, and the minimum number of matches that needs to be played in order to yield useful predictions.

An experimental evaluation on four seasons in seven different leagues shows that the proportion of correctly predicted relative positions and the mean squared error in terms of positions are natural metrics to assess the quality of the predicted final league tables. Furthermore, the evaluation shows that the traditional Elo-rating system performs well in most settings. The Pi-rating model performs reasonably well in many settings but excels when only match results from previous seasons are available.

# 2 Predictive Models

We consider multiple variants of both the widely used Elo-ratings [4, 5] and the more recent pi-ratings [2]. While these ranking systems are very similar in spirit, they differ in several important aspects. We now discuss the specifics of both ranking systems.

## 2.1 Elo Ratings

The Elo system uses a single number or *rating* to represent the strength of a team at a particular point in time. This rating increases or decreases based on the outcomes of matches. After each match, rating points are transferred from the losing team to the winning team. The number of transferred rating points depends on the difference between the ratings of the teams prior to the match. More rating points are transferred when a low-ranked team wins against a high-ranked team than when a high-ranked team beats a low-ranked team. Hence, the Elo system is self-correcting in the sense that underrated teams will gain rating points until their rating reflects their true strength, while overrated teams will loose rating points.

More formally, the rating of a team is computed as follows after each match:

$$R_{new} = R_{cur} + I \times G \times (R_{act} - R_{exp}) \tag{1}$$

In this equation, $R_{cur}$ and $R_{new}$ denote the rating of a team before and after a match, respectively. The parameter $I$ is a positive number that denotes the importance of the match, and higher numbers correspond to more important matches. The parameter $G$ is a positive number that denotes the importance of the goal difference and a bigger goal difference corresponds to an higher number. This parameter is typically defined as a function of the goal difference:

$$G = \begin{cases} 1 & \text{if GD } \leq 1 \\ 1.5 & \text{if GD } = 2 \\ \frac{(GD+11)}{8} & \text{if GD } \geq 3 \end{cases} \tag{2}$$

The parameters $R_{act}$ and $R_{exp}$ represent the actual and expected outcome of the match, respectively. The parameter $R_{act}$ takes one of three possible values: 1 for a win, 0.5 for a draw, and 0 for a loss. The parameter $R_{exp}$ takes a value between 0 and 1 and is computed as follows:

$$R_{exp} = \frac{1}{1 + 10^{\left(\frac{R_{away} - R_{home}}{400}\right)}} \tag{3}$$

In this equation, $R_{home}$ and $R_{away}$ denote the ratings of the home and away team.

## 2.2 Probabilistic Intelligence Ratings

The Probabilistic Intelligence system uses two numbers or *pi-ratings* to represent the strength of a team at a particular point in time. These numbers represent the expected goal difference against an average opponent in a league both at home and on the road. Similar to the Elo system, the ratings increase or decrease based on the outcomes of matches. After each match, the ratings of the teams involved are adjusted based on the goal difference. The key idea is that the system updates the ratings to reduce the discrepancy between the predicted and observed goal differences. The convergence speed depends on two learning rates, which denote how important recent results are for assessing the current ability of a team.

More formally, the ratings of a team are computed as follows after each match:

$$R_{\alpha,H} = C_{\alpha,H} + \psi_H(e) \times \lambda \tag{4}$$

$$R_{\alpha,A} = C_{\alpha,A} + (R_{\alpha,H} - C_{\alpha,H}) \times \gamma \tag{5}$$

$$R_{\beta,A} = C_{\beta,A} + \psi_A(e) \times \lambda \tag{6}$$

$$R_{\beta,H} = C_{\beta,H} + (R_{\beta,A} - C_{\beta,A}) \times \gamma \tag{7}$$

In these equations, $C_{\alpha,H}$ and $C_{\alpha,A}$ are the current home and away ratings for team $\alpha$, $C_{\beta,H}$ and $C_{\beta,A}$ are the current home and away ratings for team $\beta$. $R_{\alpha,H}$, $R_{\alpha,A}$, $R_{\beta,H}$, and $R_{\beta,A}$ are the respective revised ratings. Furthermore, $\lambda$ and $\gamma$ are learning rates, and $\psi(e)$ is a function of the difference between the observed and predicted goal difference, which is computed as follows:

$$\psi(e) = 3 \times \log_{10}(1 + e) \tag{8}$$

Due to space limitations, we refer to [2] for the technical details and a concrete example.

# 3 Experimental Evaluation

This section presents the experimental evaluation. We first present the dataset, the methodology and the predictive models before discussing the experimental results.

## 3.1 Dataset

In our experimental evaluation, we use a large dataset of historical football match results.[1] The dataset contains match results for eleven European football leagues including the English, German, Spanish, Italian and French top divisions. The dataset dates back to the 1993/1994 season for most leagues and contains nearly 150,000 match results. Due to space limitations, we restrict the evaluation to the four seasons from 2010 through 2014 and the following seven leagues: Belgium, England, France, Germany, Italy, The Netherlands, and Spain.

## 3.2 Methodology

To evaluate each of the models, we split the match results for each league in two sets: a training set and a test set. We use the training set to learn the parameters of the model and the test set to assess the performance of the model. We predict the outcomes of the matches in the test set in an iterative way. We use the model learned on the training set to predict the outcomes of the matches on the first match day. We then update the parameters of the model using the expected outcomes and predict the outcomes of the matches on the second match day. We repeat this procedure until all matches have been predicted.

To account for the fact that performances tend to vary throughout a season, we employ a probabilistic approach to predict match outcomes. For the Elo-rating models, we predict a match outcome by first computing the probability distribution over the possible outcomes and then sampling an outcome from this distribution. For the pi-rating model, we sample a match

[1] http://www.football-data.co.uk/data.php

outcome from a normal distribution whose mean is the expected outcome. We repeat each experiment 10,000 times and report the average results across all runs.

In each league, we predict the outcomes for the 2013/2014 season and use the match results from the earlier seasons to learn the parameters of the models.

## 3.3 Models

We compare the following four models in our experimental evaluation:

- **Random model (RM):** This model predicts a random outcome for each match.
- **Elo-rating model (EM):** This is a traditional implementation of the Elo-rating system (see Section 2.1), which represents each team by a single rating. We account for the home-field advantage by increasing the rating of the home team with an empirically determined number of rating points.
- **Split Elo-rating model (SEM):** This is a modified implementation of the Elo-rating system, which represents each team by two ratings. The first rating represents the strength at home, while the second rating represents the strength on the road.
- **Pi-rating model (PM):** This is a traditional implementation of the pi-rating system (see Section 2.2), which represents each team by two ratings.

We have not rigorously tuned the parameters. For each model, we used reasonable parameter values that work well on our large volume of training matches.

## 3.4 Results

In our evaluation, we investigate what is the most natural metric to assess a predicted league table (Q1), how well each model performs on predicting the final league table (Q2), and what proportion of the matches needs to be played in order to yield useful predictions (Q3).

#### Q1: What is the most natural metric to assess a predicted league table?

We investigated the following four metrics to assess the models:

- **Proportion of correct absolute positions:** This is the ratio between the number of correctly predicted absolute positions and the total number of positions in the league table.
- **Proportion of correct relative positions:** This is the ratio between the number of correctly predicted relative positions and the total number of relative positions in the league table. This metric compares the relative positions between any two teams.
- **Mean squared error in terms of positions:** This is the mean squared error between the actual positions and the predicted positions for each team.
- **Mean squared error in terms of points:** This is the mean squared error between the actual number of points and the predicted number of points for each team.

We found that the proportion of correctly predicted relative positions and the mean squared error in terms of positions are the two most natural metrics to assess a predicted league table. Predicting the absolute position for a team is hard because it depends on the positions of all other teams in the league. Furthermore, both the Elo-rating models and the Pi-rating model have difficulties to predict draws, which often leads to an higher variance on the predicted number of points than on the actual number of points for each team.

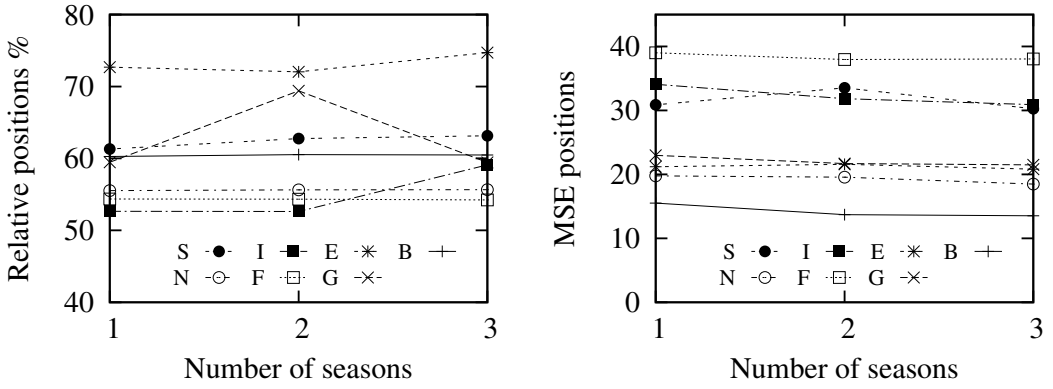


Figure 1: Learning curves for the proportion of correctly predicted relative positions and the mean squared error in terms of positions, where we vary the size of the training set.

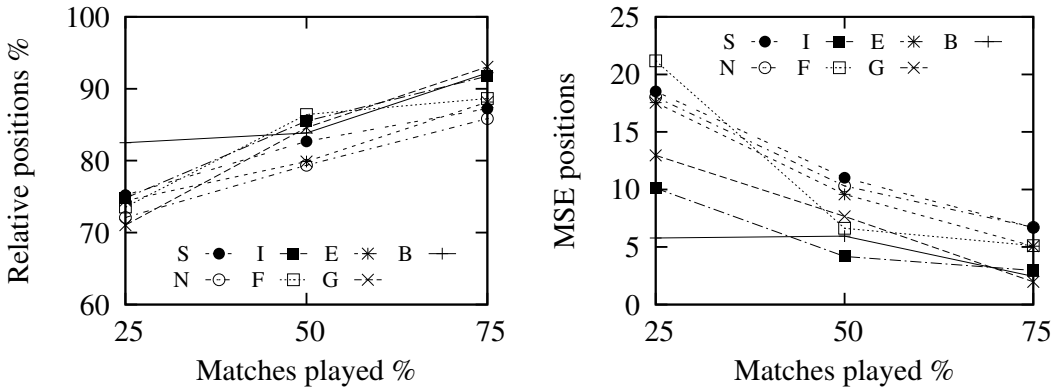


Figure 2: Learning curves for the proportion of correctly predicted relative positions and the mean squared error in terms of positions, where we vary the amount of played matches.

**Q2: How well does each model perform on predicting the final league table?**

Figure 1 shows learning curves for the proportion of correctly predicted relative positions and the mean squared error in terms of positions, where we vary the size of the training set from one season to three seasons. For the former metric, the PM performs best in 18 of the 21 settings and the SEM in the remaining three settings. For the latter metric, the EM performs best in 17 of the 21 settings, the PM in three settings, and the SEM in one setting.

**Q3: What proportion of the matches needs to be played to yield useful predictions?**

Figure 2 shows learning curves for the proportion of correctly predicted relative positions and the mean squared error in terms of positions, where we vary the number of matches played in the final season. For example, 25% means that the model is trained on the results for the 2012/2013 season as well as the results for the first quarter of the 2013/2014 season. Unsurprisingly, the

predicted final league tables become more accurate as more matches have been played. For the former metric, the EM performs best in 15 of the 21 settings and both the SEM and PM in three settings. For the latter metric, the EM performs best in 18 of the 21 settings and the SEM in the remaining three settings.

# 4 Conclusions

This paper investigates whether popular predictive rankings are capable of accurately predicting the final league tables in football leagues. The experimental evaluation shows that the proportion of correctly predicted relative positions and the mean squared error in terms of positions are the most natural metrics to assess predicted league tables. Furthermore, the evaluation shows that the traditional Elo-rating system performs particularly well in most settings we considered. While the Pi-rating model performs reasonably well in many settings as well, it excels when only match results from previous seasons are available.

The suggested directions for future research include devising more sophisticated metrics to assess the predicted final league tables (e.g., assigning weights to positions such that positions qualifying for European football become more important) and comparing to more advanced predictive ranking systems (e.g., systems that also consider performance data).

# Acknowledgments

Jan Van Haaren is supported by the Agency for Innovation by Science and Technology in Flanders (IWT). Jesse Davis is partially supported by the Research Fund KU Leuven (OT/11/051), EU FP7 Marie Curie Career Integration Grant (#294068) and FWO-Vlaanderen (G.0356.12).

# References

[1] C. Anderson and D. Sally. *The numbers game: Why everything you know about football is wrong.* Penguin UK, 2013.

[2] A. C. Constantinou and N. E. 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(1):37–50, 2013.

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

[4] A. E. Elo. *The rating of chessplayers, past and present.* Batsford London, 1978.

[5] L. M. Hvattum and H. Arntzen. Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, 26(3):460–470, 2010.

[6] D. Karlis and I. Ntzoufras. Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381–393, 2003.

[7] A. J. Lee. Modeling scores in the Premier League: Is Manchester United really the best? *Chance*, 10(1):15–19, 1997.

[8] M. J. Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982.

[9] H. Rue and O. Salvesen. Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3):399–418, 2000.