Score-based soccer match outcome modeling – an experimental review

Ondřej Hubáček*, Gustav Šourek, and Filip Železný

Czech Technical University in Prague

Abstract

In this experimental work, we propose to investigate the state-of-the-art in score-based soccer match outcome prediction modeling to identify the top-performing methods across the diverse classes of existing approaches to the problem. Namely, we bring together statistical methods based on Poisson distribution, a general ranking algorithm (Elo), domain-specific rating system (pi-ratings) and a graph-based approach to the problem (PageRank). We experimentally compare these diverse competitors altogether on a large database of soccer results to identify the true leaders in the domain.

1 Introduction

Soccer, being arguably the most popular sport in the world, continues to attract researches and practitioners competing for the design of the most accurate game result forecasting models. Indeed, there has been a plethora of such models published in the past 20 years. However, due to a lack of a standardized dataset, it has been difficult to draw conclusive statements about relative performance of the diverse approaches.

In order to gain more advantage, many of the works utilized detailed granularity of match and background information. For instance, in the top European leagues, a complete information about the game, including player-tracking data, can be obtained. However, such data are often proprietary and rather expensive, rendering them incompatible for use in academic benchmarks. Moreover, such an approach does not generalize onto the vast amount of the lower leagues, where merely the results with basic metadata is all that is being stored for each match.

To target the widest possible scope of the domain, we intersect the input information to the most common subset containing merely the match results. Such a score-based modelling paradigm allows us to predict virtually all possible matches and, consequently, unify the training and evaluation protocol across the diverse approaches.

Conveniently, a large dataset containing 218 916 match results from 52 leagues since the season 2000/01 was released recently by Dubitzky et al., 2019. The records in the dataset consist merely of the league names, dates, team names and the resulting scores. The availability of such a large dataset provides an ideal opportunity to finally shed some light onto the relative performance of the respective score-based state-of-the-art methods. For that purpose, we have started with reimplementation of the most promising models to analyze their performance under a unified protocol.

The rest of the paper is organized as follows. In Section 2 we summarize relevant research, Section 3 provides a brief description of implemented models, Section 4 explains fitting and evaluating the models, preliminary results are compiled in Section 5, conclusions and next steps can be found in Section 6.

[*] Corresponding author's email: hubacon2@fel.cvut.cz

2 Related Work

The research in the domain of score-based soccer modelling has traditionally been dominated by statistical approaches. In his pioneering work, Maher (1982) came up with a double Poison model and bivariate Poisson model. The bivariate Poisson model provided a better fit for the data. Maher also introduced the notion of teams' attacking and defensive strengths and how to use them for forecasting of the match results. This notion is still used in the current research nowadays. Dixon and Coles (1997) extended Maher's ideas, as he introduced a dependency between home and away teams' goals scored for the double Poisson model, increasing the probabilities of low-scoring draws. While Maher considered the strength of the team to be time invariant, here the idea of likelihood weighting while fitting the model was introduced. Particularly, the authors used exponential time weighting to discount the effects of past results. A different approach to the time evolution was used in Rue and Salvesen (2000), where the authors used a brownian motion to tie together the teams' strength parameters in consecutive rounds. Karlis and Ntzoufras (2003) noticed, that the bivariate Possion models tend to underestimate the probabilities of draws and introduced a diagonal-inflated bivariate Possion model. Karlis and Ntzoufras (2008) eliminated the need to explicitly model the scores dependency via utilization of Skellam distribution. The evolution of the teams' strengths was implemented using Bayesian updates. A static hierarchical model based on double Poisson distribution was introduced in Baio and Blangiardo (2010), claiming performance not inferior to the bivariate Possion model (Karlis and Ntzoufras, 2003). Koopman and Lit (2015) introduced time dynamics into the bivariate Poisson model using a state space model representation. Authors pointed out that the dependency between scores had a little effect on out-of-sample forecasting performance of the model. Angelini and De Angelis (2017) investigated another technique for implementing the time-dynamics with a PARX model (Agosto et al., 2016). The PARX model outperformed Dixon and Coles (1997) in forecasting number of scored goals.

The most recent novelty in statistical approaches is the use of bivariate Weibull count model (Boshnakov et al., 2017). Unlike in the Poisson distribution, where the mean is equal to the variance, the Weibull count distribution is determined by two parameters, allowing for better handling of both under and over dispersed data. The bivariate model is constructed using a copula function. The model provides a better fit for the data than the Poisson model at the expense of a higher computational time, as the computation of the probability density function of the Weibull count model is computationally demanding. A great review of the statistical approaches can be found in Ley et al. (2019).

Another technique to estimate the strength of an individual or a team are the so-called rating systems. The world's best known rating system is the Elo rating (Elo, 1978), originally used for assessing the strength of chess players. The player's performance is assumed to be drawn from a Gaussian distribution with fixed variance. The mean of such distribution is then the player's rating (skill). An application of Elo rating in the domain of soccer was shown in Hvattum and Arntzen (2010). While the authors have not provided a sufficient comparison against other models, a recent work by Robberechts and Davis (2018) demonstrated that the method is sound. Trueskill (Graepel et al., 2007) enhances the Elo rating as it operates not only with the variance of the player's performance but also with the variance of his skill (rating). This variance reflects the uncertainty about player's skills, when we have not observed enough data (performances). The author demonstrated faster convergence and better predictive performance in comparison with the Elo rating. One of the caveats of the Truskill is that in does not propagate the newly obtained information backward to correct the ratings. In other words, it does filtering instead of smoothing. The work by Dangauthier et al. (2008) aimed to fix this issue. Also, the plain version of Truskill does not account for the score difference, as it only considers the win-draw-loss outcome of a match. Guo et al., 2012 proposed an extension to handle the score differences and claimed superior performance to the vanilla Trueskill, also on a soccer dataset. The current evolution of the Trueskill rating system is Trueskill2 (Minka et al., 2018), however most of the improvements are domain specific to matchmaking in online games, which is the primary focus of the system. A soccer domain-specific rating system called pi-ratings was introduced in Constantinou and Fenton (2013). The team's strength is represented by its' home and away ratings, that are updated after each match according to manually set learning rates. Another score-based rating system was developed by Berrar et al. (2019). The rating system parameters were tuned using particle swarm optimization and fed to a standard off-the-shelf learner.

Machine learning models are not very common in score-based modelling as they usually leverage on extra features besides the scores or ratings. Some recent exceptions were the models for the 2017 Soccer Prediction Challenge (Dubitzky et al., 2019), where the dataset contained merely the historical results with basic metadata on the matches. For the challenge, Constantinou (2019) extended his pi-ratings model with a Bayesian network to obtain the probability distribution over possible match outcomes from the rating difference. Tsokos et al. (2019) tested several variations of Bradley-Terry model and hierarchical Possion model. In the end, the hierarchical Possion model outperformed all the Bradley-Terry models. The inferiority of Bradley-Terry model to other methods was further confirmed by Ley et al. (2019).

The relational structure of the data was pointed out by Van Haaren and Van den Broeck (2015) where the authors achieved promising results. An advanced relational learner (Natarajan et al., 2012) was also tested in Hubáček et al. (2019), however with a little success. The same authors later proposed relational team embeddings (Hubáček et al., 2018), implemented in a framework for combining relational and neural learning (Sourek et al., 2018), with more promising results. The graph representation of the data was also utilized by Govan, Meyer, et al. (2008), who used the PageRank (Page et al., 1999) to estimate the teams' strengths. The same author later proposed a so-called offense-defense model (Govan, Langville, et al., 2009), that can be seen as an analogy to the HITS algorithm (Kleinberg, 1999).

3 Models

In this section, we introduce the models we have reimplemented and tested so far. The selected models are considered to be very competitive in their respective categories. The Double Poisson model proved to be very competitive in the recent comparison of statistical models (Ley et al., 2019). Robberechts and Davis (2018) demonstrated effectiveness of the Elo ratings, while Constantinou and Fenton (2013) proposed their improvement – the pi-ratings. The PageRank model represents the category of models that utilize the graph structure of the data. This paper presents a work in progress, and we plan to broaden the portfolio of tested models further.

3.1 Double Poisson Model

Double Poisson model represents one of the earliest (Maher, 1982) and simplest models. However, as was shown in Ley et al. (2019), it is still very competitive nowadays. The model assumes the goals scored by the competing teams in a match to be independent. Therefore, the probability of a home team scoring x goals with the away team scoring y goals is given by

$$P(G_H = x, G_A = y | \lambda_H, \lambda_A) = \frac{\lambda_H^x e^{-\lambda_H}}{x!} \cdot \frac{\lambda_A^y e^{-\lambda_A}}{y!},$$

where λ_H and λ_A are the scoring rates of the teams (the means of the underlying Possion distributions). The scoring rates for a match for the teams can be expressed in terms of Maher's specification as

$$log(\lambda_H) = Att_H - Def_A + H$$
$$log(\lambda_A) = Att_A - Def_H$$

where H represents a home advantage, and Att and Def are respectively the defensive and offensive strengths of the teams (the actual model parameters).

Later, Ley et al. (2019) demonstrated that the number of the model's parameters can be effectively halved by considering only a single strength parameter for each team without any loss of predictive performance, i.e. reducing to

$$log(\lambda_H) = Str_H - Str_A + H$$
$$log(\lambda_A) = Str_A - Str_H$$

3.2 Elo Ratings

The Elo (Elo, 1978) is a general rating system the modification of which is still used for evaluation of the strength of chess players. Hvattum and Arntzen (2010) proposed its modification for soccer and consequently Robberechts and Davis (2018) demonstrated effectiveness of the method. The modification involves the use of an ordered logit model (McCullagh, 1980) to obtain the probability distribution over the possible match outcomes. At the core, each the team's performance is assumed to be normally distributed around its true strength. The expected scores for both teams are then calculated as follows

$$E^H = \frac{1}{1 + c^{(R^A - R^H)/d}}$$
$$E^A = 1 - E^H$$

where R^H and R^A are the ratings of the home and away teams, and c and d are metaparameters of the model. The actual outcome of the match is then numerically encoded as

$$S^H = \begin{cases} 1 & \text{if the home team won} \\ 0.5 & \text{if the match was drawn} \\ 0 & \text{if the home team lost} \end{cases}$$

Finally after the match, the ratings of both the teams are updated w.r.t.

$$R^H_{t+1} = R^H_t + k(1 + \delta)^\gamma \cdot (S^H - E^H)$$

$$R^A_{t+1} = R^A_t - k(1 + \delta)^\gamma \cdot (S^H - E^H)$$

where δ is an absolute goal difference, k represent a learning rate and γ is a metaparameter scaling the influence of the goal difference on the rating change.

3.3 pi-ratings

The pi-ratings (Constantinou and Fenton, 2013) represent a domain-specific rating system. Each team is assigned two ratings, representing it's strength when playing home and when playing away. For each match, expected goal difference is calculated, based on home team's *home rating* and away's team *away rating* . After the match is played, the *expected score* is compared to the actual outcome. If a team performs better than expected, its ratings are increased based on the discrepancy of the actual outcome and expected outcome and the learning rates (metaparameters of the model). Both team's home and away ratings get updated after a match, with both updates having a separate learning rate. We refer to the original paper for more details (Constantinou and Fenton, 2013). Finally, the probability distribution over the possible mach outcomes is once again determined by an ordered logit model.

3.4 PageRank

The PageRank (Page et al., 1999) algorithm was originally designed for assessing importance of web pages. In the original algorithm, the directed edge (p_i, p_j) represents a link from page p_i to p_j. The importance of a webpage is proportional to the probability of a random walk over the webgraph visiting the page. As was shown by Govan, Meyer, et al. (2008) it can be similarly used for ranking of teams in a competition. The competition can be represented as a graph, where the nodes represent the teams and the edges represent the matches between them. For our model, the adjacency matrix M as defined as follows:

$$M_{ij} = \frac{\sum_m PTS_j(m) \cdot w_m}{\sum_m w_m}$$

where $PTS_j(m)$ is the number of points team j got from match m against team i and w_m is the weight of the match. This model represents a weighted version of the PageRank used by Hubáček et al. (2019).

4 Validation Framework

All the data used in this review came from the Open International Soccer Database v2 (Dubitzky et al., 2019). We limited the original database to seasons ranging from 2000/01 to 2005/06 to prevent data contamination in future experiments. Still, this subset provided us with nearly 60 000 of matches from 38 leagues and 27 countries. The first season of each league was only used as a warm-up season, omitted from model evaluation. Furthermore, the first 5 rounds of each season were also used as a burn-in period, omitted from the evaluation, too. We evaluated the models using ranked probability score (Epstein, 1969) and accuracy.

4.1 Model fitting

For fitting of models' free parameters we used common optimization routine based on the L-BFGS-B algorithm (Byrd et al., 1995). The fitting process and hyperparameter settings for each of the selected models is specified bellow.

Double Poisson Model Model's parameters are found by maximizing the weighted likelihood of the observed data

$$L = \prod (P(G_i^H = x, G_i^A = y | \lambda_i^H, \lambda_i^A) \cdot w_i)$$

Table 1: Comparison of the RPS and Accuracy of the tested models.

	RPS	Accuracy
Double Poisson	0.2082	0.4888
Elo	0.2088	0.4887
pi-ratings	0.2092	0.4897
PageRank	0.2128	0.4775

where w_i represents the weight of each observation. Since the first successful application (Dixon and Coles, 1997), exponential time weighting is being commonly used as

$$w_i = e^{-\alpha t_\Delta}$$

where t_Δ is the time passed since the match was played and α is a metaparameter. We use $\alpha = 0.0019$ as was done in Ley et al. (2019). The parameters are refitted after each league round to account for the newly obtained information.

Elo & pi-ratings Elo ratings and pi-ratings require 2 and 3 metaparameters respectively, and 3 parameters for the subsequent ordered logit model. These parameters are optimized jointly, minimizing the average RPS on previous seasons. The ratings are updated after each league round, while the (meta)parameters are refitted after each season.

PageRank The PageRank requires a setting of 1 metaparameter – the damping factor (= 0.25), which was tuned manually. The 3 parameters of the subsequent ordered logit model are optimized by minimizing the average RPS on previous seasons. The weight of each match is computed in the same way as in the double Poisson model. The ratings are recalculated after each league round. The parameters of the ordered logit model are refitted after each season.

5 Preliminary Results

The result are summarized in Table 1. The double Possion model outperformed all the models in terms of RPS. The pi-ratings had a marginally higher accuracy. The PageRank trailed significantly behind other tested models both in RPS and accuracy.

The inferiority of the PageRank model could have its base in the fact that the other models leverage the scores of the teams, while the PageRank utilizes only the win/draw/loss outcome of the match. Here, we proposed a weighted version of the PageRank algorithm, which performed better than the original unweighted version (RPS of 0.2140). There are still countless ways how to advance construction of the adjacency matrix for the PageRank approach. For instance, integrating the scores into the adjacency matrix could lead to further improvements.

6 Conclusion

In this work we compared performance of diverse models for predicting soccer match outcomes based solely on historical results. Double Poisson model, one of the very oldest models in soccer forecasting, performed the best in terms of RPS. Pi-ratings, the newest model from our comparison, on the other hand outperformed the remaining models in terms of predictive accuracy. While it was previously shown that the double Possion model is, despite it's simplicity, competitive among other statistical models (Ley et al., 2019), we can see it holds its ground against more diverse competitors as well.

Future work The work described in this paper is still in progress, and we plan to further extend on this review in various directions. Most importantly, we have merely compared 4 models so far, however we intend to update the portfolio of methods towards an extensive comparison of state-of-the-art in the domain. Regarding optimization of the models tested, we have optimized the metaparameters of the Elo and pi-ratings jointly with the parameters of subsequent ordered logit model, as was done by Robberechts and Davis (2018). Here we further plan to try out also a 2-step optimization protocol, where the optimizations of metaparameters and parameters are handled by two different optimizers. Moreover, we will investigate the influence of using a multinomial regression instead of the ordered logit model, as well as using more rating features as input covariates. Finally, with the complete set of models and optimization routines, we will extend our dataset to the full scope of available data.

References

Agosto, Arianna et al. (2016). "Modeling corporate defaults: Poisson autoregressions with exogenous covariates (PARX)". In: *Journal of Empirical Finance* 38, pp. 640–663.

Angelini, Giovanni and Luca De Angelis (2017). "PARX model for football match predictions". In: *Journal of Forecasting* 36.7, pp. 795–807.

Baio, Gianluca and Marta Blangiardo (2010). "Bayesian hierarchical model for the prediction of football results". In: *Journal of Applied Statistics* 37.2, pp. 253–264.

Berrar, Daniel, Philippe Lopes, and Werner Dubitzky (2019). "Incorporating domain knowledge in machine learning for soccer outcome prediction". In: *Machine Learning* 108.1, pp. 97–126.

Boshnakov, Georgi, Tarak Kharrat, and Ian G McHale (2017). "A bivariate Weibull count model for forecasting association football scores". In: *International Journal of Forecasting* 33.2, pp. 458–466.

Byrd, Richard H et al. (1995). "A limited memory algorithm for bound constrained optimization". In: *SIAM Journal on Scientific Computing* 16.5, pp. 1190–1208.

Constantinou, Anthony C (2019). "Dolores: a model that predicts football match outcomes from all over the world". In: *Machine Learning* 108.1, pp. 49–75.

Constantinou, Anthony C and Norman Elliott Fenton (2013). "Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries". In: *Journal of Quantitative Analysis in Sports* 9.1, pp. 37–50.

Dangauthier, Pierre et al. (2008). "Trueskill through time: Revisiting the history of chess". In: *Advances in neural information processing systems*, pp. 337–344.

Dixon, Mark J and Stuart G Coles (1997). "Modelling association football scores and inefficiencies in the football betting market". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 46.2, pp. 265–280.

Dubitzky, Werner et al. (2019). "The Open International Soccer Database for machine learning". In: *Machine Learning* 108.1, pp. 9–28.

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

Epstein, Edward S (1969). "A scoring system for probability forecasts of ranked categories". In: *Journal of Applied Meteorology* 8.6, pp. 985–987.

Govan, Anjela Y, Amy N Langville, and Carl D Meyer (2009). "Offense-defense approach to ranking team sports". In: *Journal of Quantitative Analysis in Sports* 5.1.

Govan, Anjela Y, Carl D Meyer, and Russell Albright (2008). "Generalizing Google's PageRank to rank national football league teams". In: *Proceedings of the SAS Global Forum*. Vol. 2008.

Graepel, Thore, Tom Minka, and R TrueSkill Herbrich (2007). "A Bayesian skill rating system". In: *Advances in Neural Information Processing Systems* 19, pp. 569–576.

Guo, Shengbo et al. (2012). "Score-based bayesian skill learning". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 106–121.

Hubáček, Ondřej, Gustav Šourek, and Filip Železný (2018). "Lifted Relational Team Embeddings for Predictive Sport Analytics". In: *Proceedings of the 28th International Conference on Inductive Logic Programming*. CEUR-WS.org, pp. 84–91.

Hubáček, Ondřej, Gustav Šourek, and Filip Železný (2019). "Learning to predict soccer results from relational data with gradient boosted trees". In: *Machine Learning* 108.1, pp. 29–47.

Hvattum, Lars Magnus and Halvard Arntzen (2010). "Using ELO ratings for match result prediction in association football". In: *International Journal of forecasting* 26.3, pp. 460–470.

Karlis, Dimitris and Ioannis Ntzoufras (2003). "Analysis of sports data by using bivariate Poisson models". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 52.3, pp. 381–393.

Karlis, Dimitris and Ioannis Ntzoufras (2008). "Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference". In: *IMA Journal of Management Mathematics* 20.2, pp. 133–145.

Kleinberg, Jon M (1999). "Authoritative sources in a hyperlinked environment". In: *Journal of the ACM (JACM)* 46.5, pp. 604–632.

Koopman, Siem Jan and Rutger Lit (2015). "A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178.1, pp. 167–186.

Ley, Christophe, Tom Van de Wiele, and Hans Van Eetvelde (2019). "Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches". In: *Statistical Modelling* 19.1, pp. 55–77.

Maher, Michael J (1982). "Modelling association football scores". In: *Statistica Neerlandica* 36.3, pp. 109–118.

McCullagh, Peter (1980). "Regression models for ordinal data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 42.2, pp. 109–127.

Minka, Tom, Ryan Cleven, and Yordan Zaykov (2018). "TrueSkill 2: An improved Bayesian skill rating system". In: *Tech. Rep.*

Natarajan, Sriraam et al. (2012). "Gradient-based boosting for statistical relational learning: The relational dependency network case". In: *Machine Learning* 86.1, pp. 25–56.

Page, Lawrence et al. (1999). *The PageRank citation ranking: Bringing order to the web.* Tech. rep. Stanford InfoLab.

Robberechts, Pieter and Jesse Davis (2018). "Forecasting the FIFA World Cup–Combining result-and goal-based team ability parameters". In: *Machine Learning and Data Mining for Sports Analytics ECML/PKDD 2018 workshop*. Vol. 2284. Springer, pp. 52–66.

Rue, Havard and Oyvind Salvesen (2000). "Prediction and retrospective analysis of soccer matches in a league". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 49.3, pp. 399–418.

Sourek, Gustav et al. (2018). "Lifted relational neural networks: Efficient learning of latent relational structures". In: *Journal of Artificial Intelligence Research* 62, pp. 69–100.

Tsokos, Alkeos et al. (2019). "Modeling outcomes of soccer matches". In: *Machine Learning* 108.1, pp. 77–95.

Van Haaren, Jan and Guy Van den Broeck (2015). “Relational learning for football-related predictions”. In: *Latest Advances in Inductive Logic Programming.* World Scientific, pp. 237–244.