Wharton University of Pennsylvania Sports Analytics and Business Initiative

2024, Fall

wsb.wharton.upenn.edu/wharton-sports-analytics-journal-2

Calculating Win Probabilities of Any Matchup of Soccer Teams: A Whole-History Rating Approach for the Wharton High School Data Science Competition

Ethan Seung, *'24*
Harvard-Westlake School, CA, USA

John Xu, *'24*
Harvard-Westlake School, CA, USA

Ryder Katz, *'24*
Harvard-Westlake School, CA, USA

Mason Wetzstein, *'24*
Harvard-Westlake School, CA, USA

Michael Barr, *'24*
Harvard-Westlake School, CA, USA

Advisor: Andy Stout
Harvard-Westlake School, CA, USA

2024 Wharton High School Data Science Competition
1st Place Team

Abstract

This study presents a novel approach to evaluating soccer team strength and calculating win probabilities given any two matchups of soccer teams in the fictional North American Soccer League (NSL), whose data is provided for the Wharton High School Data Science Competition. Traditional win-loss metrics have many limitations, like failing to account for the strength of schedule, margin of victory, and home-field advantage. Thus, we developed a Whole-History Rating (WHR) model adapted from Rémi Coulom that incorporates expected goals (xG) and accounts for home-field advantage, overcoming the sequencing issues inherent in single-pass Elo systems. Using a dataset of 476 NSL games, we implemented a multi-pass system for iterative rating adjustments and employed Bayesian inference to develop probability distributions for team ratings.

The model was optimized using Nelder-Mead optimization, achieving a root mean square error (RMSE) of 0.346 when tested against a held-out dataset. Key findings include a significant home-field advantage constant of 32.3 rating points and identifying the top four teams based on the optimized ratings. The model's strengths lie in its order-independent nature, use of expected goals as a performance metric and continuous updating of all team ratings. Limitations and areas for future work include refining the expected goals metric to account for elite player performance and developing a dynamic home-field advantage factor that considers time zone differences. This research provides a more nuanced and accurate method for evaluating team strength and forecasting game outcomes in soccer leagues around the world.

key words: soccer; win probability; xG; home-field advantage; ELO

Introduction

The Wharton High School Data Science Competition presented a unique challenge: accurately evaluating soccer team strength given 476 games of data in the fictional North American Soccer League (NSL). This scenario is similar to real-world challenges in professional soccer, where accurately assessing team capabilities is crucial for various stakeholders, including team management, fans, and analysts. A notable difference, however, in this fictional league was that each team's strength of performance remained constant throughout the season, meaning that teams did not improve or were affected by player injuries.

Our task was twofold:

1. Determine individual match outcomes in the group stage of the NSL playoffs, where:
    a. 28 teams are divided into seven groups of four teams
    b. Within each group, each of the four teams plays three games
    c. A total of six games per group and 42 games for the entire group round
    d. Home field advantage exists in this stage

2. Calculate win percentage probabilities in the knockout round, where:
    a. Four teams move on to the Knockout Round
    b. All Knockout Round games are played at a neutral site (no home-field advantage)

The traditional metric for evaluating team performance by win-loss record has numerous flaws. It fails to account for the quality of opposition (strength of schedule) and margin of victory (strength of play). Goal differential is commonly used to fix this, but it can be heavily influenced by random factors or "lucky" outcomes, such as deflections, goalkeeper errors, or unusual weather conditions. Thus, goal differential may not accurately reflect a team's strength over a larger sample size, while also being skewed by in-game score-based strategy changes. Additionally, the traditional metric fails to account for the variance introduced by luck and random chance in outcomes. For example, an unlucky bounce of the ball is all it takes to swing the outcome of a game.

Finally, home-field advantage, particularly in this league, drastically skews win-loss records. Penalty kicks were awarded to home teams at a disproportionate ratio of 126:53. We found a strong negative correlation with distance traveled to match and match results. Home-field advantage was prevalent in regular-season play and group-stage games.

In the controlled environment of the fictional NSL, we set out to go beyond simple methods of team evaluation, and develop a complex, precise rating system to not only predict playoff outcomes, but their probabilities.

A common way to solve this issue is through the classic Elo system, often utilized to rank chess players. It adjusts each competitor's rating points based on the outcome of their games and their opponent's strength; thus, winning against a higher-rated opponent yields more points than defeating a lower-rated one.

The classic Elo system operates on a "single-pass" basis, adjusting ratings after each match only once. It gives more weight to the results of games later in a season, thus dependent on a chronological sequence of games or a timeline. However, the dataset did not specify the games' timeline and each team's strength of performance remained constant as defined earlier, so we hypothesized that a classic Elo system would not work.

To demonstrate the issue with this chronological dependency, we created and tested a model of the classic Elo system, processing games in both forward and backward numerical order. On a base rating of 1500, we found swings of up to 430 rating points per team, showing the immense impact sequencing had on team evaluations and exposing the flaw of a single-pass system.

Methods

Rational for Methodology Selection

While the traditional, single-pass Elo system didn't work, the fundamental idea behind an Elo system to account for the strength of schedule and find true ratings of teams to calculate win probabilities is feasible with certain tweaks. Specifically, we needed to eliminate the influence of the order of games, evaluate the impact of home-field advantage, and build the system using a metric that isn't simply wins and losses. Not only did our solution need to be able to correctly predict outcomes, but the probability of those outcomes.

We created a Whole-History Rating (WHR) model, to solve the ordering issue and factored in a constant to account for home-field advantage. This model, instead of only adjusting a team's rating after their games, constantly adjusts the ratings of all teams in the league after every outcome, eliminating the sequencing issue present in single-pass Elo systems.

Though the WHR model is a great solution, it is only as useful as the metrics on which we center it. Thus, due to the flaws mentioned previously with wins and losses, we decided to use the difference in cumulative expected goals (xG) to award rating points. This reduces the influence of randomness and luck in individual shots because expected goals are derived from the historical averages of shot outcomes. In addition, since expected goals evaluate the quality of scoring chances—factoring in variables such as distance from goal, shot angle, and player skill—it can provide a detailed measure of team capabilities and match dynamics. Consequently, using expected goals as a basis for awarding rating points directly links rating

adjustments to these measurable performance indicators rather than solely to outcomes like wins and losses.

Whole-History Rating (WHR) Model

Our model utilized a dataset of 476 NSL games provided by the Wharton High School Data Science Competition. This dataset included information on:

- Match outcomes
- Expected goals for each team
- Home/away status
- Other metrics (time of possession, shots on goal, corner kicks, and penalty kicks)

The data was randomly split into a training set (80%, 380 games) and a test set (20%, 96 games) using Python. We developed the WHR model to overcome the limitations of traditional Elo systems, particularly the dependency on game order. The model consists of two main components: rating point allocation based on game outcomes, and multi-pass system for iterative rating adjustments

Rating Point Allocation

We employed the Bradley-Terry model to calculate expected win probabilities based on current team ratings. The probability of team *i* winning against team *j* is given by:

$$P(i\ beats\ j) = \frac{1}{1 + b^{\frac{(r_i - (r_j + h))}{c}}}$$

where r_i and r_j are the ratings of teams i and j respectively, h is the home field advantage constant, which gives the home team a boosted amount of rating points, c is the optimal rating-to-odds conversion—that is, fitting how much, for example, a 50-point rating difference translates to win probability, and b is the base for our Bradley-Terry model.
Rating points were then adjusted using the equation below.

$$r_i = k\left(1 + \frac{(xG_1 - xG_2)}{d}\right)(s_i - e_i)$$

This equation incorporates:

1. s_i - The expected win probability calculated using the initial rating points in the Bradley-Terry model above.
2. e_i - The actual match outcome, defined as 0 for a loss, 0.5 for a draw, or 1 for a win by team i.
3. (xG_1 - xG_2) - The difference in cumulative expected goals between the teams.

4. d - A scaling factor for the expected goals difference.
5. k - An overall adjustment factor.

The equation adjusts the rating (r_i) based on the difference between the expected probability (s_i) and the actual outcome (e_i), further weighted by the expected goals difference. This allows the model to account for both match results and the magnitude of team performance as measured by expected goals.

Multi-pass System

The model implements a multi-pass system that iteratively adjusts ratings for all teams after each match. This approach ensures that the impact of each game result propagates through the entire network of team connections. The process works as follows:

1. Initialize all teams with a base rating of 1500 points.
2. For each game in the dataset:
    a. Calculate expected win probabilities using current ratings.
    b. Adjust ratings based on the difference between expected and actual outcomes.
    c. Recalculate ratings for all teams that have played against the teams in the current game.
    d. Repeat steps a-c for all previous games involving the affected teams.

Bayesian Inference

Using Bayesian Inference, we inferred the probability that a given team rating causes the outcome of all games in the model so far, thus creating a distribution. Each team's probability distribution contains the likelihood of any given rating points being the actual strength of that team based on all actual game results passed through the model so far.

When a new game is passed into the model, the probability distribution of each team is recalculated according to Bayes' Theorem, relying on win probabilities from the Bradley-Terry model to establish a prior probability distribution for each team's ratings:

$$P(\textit{Rating} \mid \textit{Game Result}) = P(\textit{Game Result} \mid \textit{Rating}) \times P(\textit{Rating})$$

In this way, all teams connected by game history to the teams in the current game, including teams that have played against them, or teams linked indirectly through competition, change their rating accordingly to reflect the newest game result.

After obtaining the posterior probability distribution for each team, we can find the rating that yields the maximum probability using Newton's method—essentially identifying the most probable rating within the distribution and, thus, the rating most likely to have caused the observed game results.

Since the most probable rating for each team is now updated, we need to recalculate the win probabilities and the rating points given for all the previous matches a given team has played in.

We do this by redoing Rating Point Allocation with the new most probable rating. This process repeats after every game is inputted into the model.

Once all games have been processed, the ratings that give the maximum of each team's final distribution curve become the official ratings for each team that we use for real match predictions.

Optimization & Validation

After creating our Whole-History Rating model, it was tested against the test set—the remaining 20% of games we had originally set aside. From this, we calculated the RMSE by comparing our predictions to the actual outcomes of those games decided by expected goals.

Then, using Nelder-Mead optimization, we found the set of variable values in our winning probability and rating point allocation equations that returned the lowest RMSE:

- Home field advantage constant
- Weighting of expected goal differences
- Rating-to-odds conversion constant
- Win probability base
- Ratings update divisor

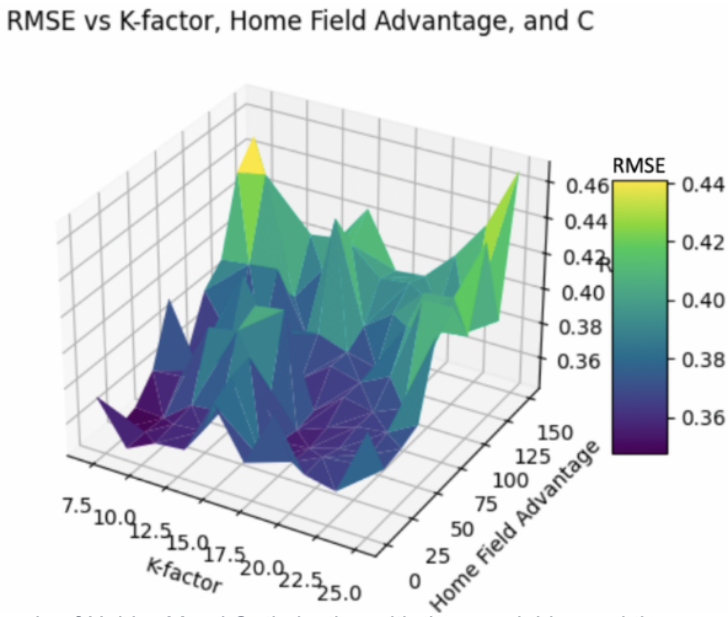

Figure 1. An example of Nelder-Mead Optimization with three variables and the corresponding RMSE.

We were able to get a resulting RMSE of .346, indicating a close fit between model predictions and actual game outcomes. We recognize that this number cannot go much lower because of randomness in the game. Finally, after finding the variables that gave us the lowest RMSE, we retrained our Whole-History Rating model on all 476 games of data to create our final ratings for group stage and knockout predictions.

Results

Optimized Parameters

the table below shows the results of our Nelder-Mead Optimization. The result of the home-field advantage constant of 32.317 means that an additional 32.317 rating points are given to the home team before any win probabilities are calculated. If two teams were 1500-rated, this would translate to a 56.5% win probability, giving a sizable advantage to the home team. This constant was used to predict the group stage games where home-field advantage was prevalent but was set to 0 for the knockout games where they were played at a neutral site.

K factor (k):	15.460
Home-field advantage constant (h)	32.317
Team ratings to odds conversion factor (c)	318.788
Win probability base (b)	11.599
Ratings update divisor (d)	2.360

Team Ratings

Final team ratings were calculated using the optimized WHR model trained on the full dataset of 476 games. The full league team rating are shown in Fig 2. The top four teams based on these ratings were:

1. Fargo Falcons
2. Anchorage Avalanche
3. Boise Thunderhawks
4. Dover Dragons

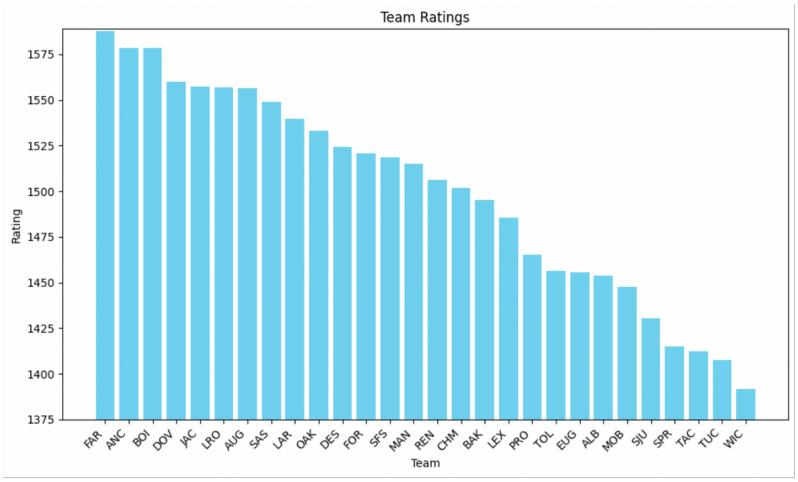

Figure 2. Final full NSL team ratings

Discussion

Model Strengths

The WHR model developed for this competition offers several advantages over traditional rating systems. The model has independence from game order, allowing for robust ratings in datasets without clear chronology. The model incorporation of expected goals as a more nuanced performance metric than wins or losses. The model enables explicit consideration of home-field advantage. Finally, the model has continuous updating of all team ratings based on new game results, this captures indirect relationships between teams.

Limitations and Future Work

More research is needed to address the following limitations related to expected goals and home field advantage.

- Expected Goals:
    - Fail to account for the ability of elite playmakers who are better shooters or goalies, as they are based on historical averages of shot outcomes in specific situations
    - Fail to quantify the effects of real game strategy, where play revolves around the actual score and not expected goals. Thus, a team may play more defensively and shoot less when winning in actual score despite being behind in expected goals.
- Home-Field Advantage Constant:
    - Generalizes the home-field advantage to be the same across all games and contexts—the 32.3 home-field advantage constant gave all home teams a 56.5% win probability against an evenly matched opponent. However, our findings show that larger time zone differences for away teams, especially above three hours, significantly increased goal differential in favor of the home team. This indicates the need for a dynamic factor in calculating home-field advantage.

Figure 3. Home team advantage by travel time for the away team (AwayHours)

Acknowledgments

We would like to thank Mr. Andy Stout, our team advisor, who supported us throughout the competition by giving us feedback on our presentation. Special thanks to the organizers of the Wharton High School Data Science Competition for providing the challenging problem and the dataset that made this research possible.

References

Assessing the performance of Premier League goalscorers. Stats Perform. (2012, April 9). https://www.statsperform.com/resource/assessing-the-performance-of-premier-league-goalscorers.

Coulom, R. (2008, September). Whole-history rating: A Bayesian rating system for players of time-varying strength. In *International conference on computers and games* (pp. 113-124). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-87608-3_11

Firth, D. (2005). Bradley-Terry Models in R. *Journal of Statistical Software*, *12*, 1–12. https://doi.org/10.18637/jss.v012.i01