SAARLAND UNIVERSITY

Data Science and Artificial Intelligence (DSAI) Project Seminar: Sports Analytics Tasks

Pascal Bauer[1] and Luis Holzhauer[1]

[1]Chair of Sports Analytics, Saarland University
pascal.bauer@uni-saarland.de/luho00006@uni-saarland.de

April 24, 2025

1 Team and Player Strength Models

Ranking systems, such as the ELO algorithm and similar methodologies like Glicko [6], are widely used in sports ranging from individual competitions like chess [3], table tennis [9], or tennis [8], to team-sports like football [5]. These systems aim to estimate player strength while enabling fair ranking systems even when only a limited subset of players compete against each other.

While official ranking systems can only be fed by pure results for fairness reasons, player and team-strength models can also integrate potentially predictive factors like players age, height, handedness, etc.

The objective of this project is to reimplement ELO ratings and official ranking systems using real-world-data and compare their predictability for further matches with own-designed player- / team-strength models in either Tennis, Football (soccer) or Cricket.

1.0.1 Task 1: ELO rating

The development of a rating system is a common task in sports. Particularly, the ELO rating system [1] is an established method. The major task will be to create and apply an ELO rankings algorithm to one of the respective sports (tennis, cricket or football/soccer).

1. Develop a function to implement the ELO Rating System with arbitrary hyperparameter K and s.

[1] `https://en.wikipedia.org/wiki/Elo_rating_system`

2. Apply the rating system to the respective dataset (depending of the sport you choose) with starting values $R_0 = 100$, $s = 15$ and $K = 15$.[2]

3. Develop an approach that finds the optimal values for s and K based on the respective season and display the final ranking table at the end of the season.

4. Develop and implement a strategy for quantifying the single most surprising win among men's match results in the respective dataset. Summarize your approach and findings.[3]

5. *Optional Task:* Implement the Glicko function [6], tune the hyperparameters and compare the results with the ELO rating.

6. Summarize your approach and findings.

1.0.2 Task 2 (*Optional Task)*: Re-implement Official Ranking Systems per Sport

Many global sports organizations (like ATP, FIFA, FIDE) have official ranking systems which are used to define draw pots and thus heavily influence tournament schedules. Although ELO is an established basis for ranking systems, some sports got their individual methods.

1. Research the official ratings in the chosen sport (typically regulated by the governing association like FIFA, FIDE) and re-implement the algorithms accordingly.

2. If different, compare the ELO- and the official rankings by checking a general correlation and explore differences. Feel free to also explore how ranking systems from certain sports would perform in other sports.

1.0.3 Task 3: Player strength model

The basic Elo rating is based on players/teams previous performances and awards points based on the opponent's Elo relative to the own Elo and the outcome of the competition. Nevertheless, based on the sport specific key performance indicators, other factors (like players age, height, handedness in tennis) influence performances as well and may significantly improve the performance of a player strength model. Therefore, the task is to develop a nuanced player strength model based on the available datasets.

1. Identify key performance indicators based on the existing literature (or define them), which can be used to build a nuanced player strength model in the respective sport.

[2] Tennis: Years 2020 until 2024 (included) — Football: Bundesliga 2015/2016 season with starting — Cricket: men's ODI match results in 2020.

[3] Tennis: Years 2020 until 2024 (included) — Football: Bundesliga 2015/2016 season with starting — Cricket: men's ODI match results in 2020.

2. Build a machine learning model which predicts future success (win/ draw/ loose or more granular metrics considering the dominance of a victory). You should explore different machine learning models, for example (logistic) regressions, tree-based models, or XGBoost models.

3. Validate the results of your player strength model with the basic Elo approach from task 1 and the official sport rankings (if different) model by using machine learning algorithms making pre-match win predictions. Make sure to validate out-of-sample and avoid overfitting. Report your results appropriately and scientifically (integrate literature, graphs, and tables as seen fit).

1.0.4 Task 4 (*Optional Task)*: Betting odds

Betting providers rely on strong player- / team-strength models in order to predict future results. Although their odds are built based on these models, they optimize them toward an optimal margin instead of towards optimal match predictions —particularly , they make use of human (seven local) biases like the overestimation of their favorite team performance or the over-representation of emotions.

1. Research available betting odds in the respective sports, e.g. via APIs [4]. Transform the betting odds into a team-strength model and future out-of-sample match outcome predictions. Betting margins and local adjustments could be normalized by averaging the odds of multiple vendors.

2. Compare the betting odds with previous approaches.

1.1 Tennis

1.1.1 Tennis dataset

Jeff Sackmann provides an extensive dataset of approximately 65.000 ATP players including attributes such as *ID, Name, Hand, Country* and *Height* in his open-source repository.[5] Additionally, match outcome data exists for over $30k$ singles matches between 1968 and 2024. This comprehensive dataset is further enriched by a substantial big point-by-point dataset of over 9000 matches during the seasons 2011-2024. Rankings and points of the ATP and WTA can be scraped from the official ATP and WTA websites using an open-source Git repository:[6] The combination of these datasets offers a robust foundation for developing and evaluating alternative player strength models.

[4] For example: https://the-odds-api.com/

[5] GitHub: `https://github.com/JeffSackmann/tennis_atp/blob/a36a13fe21f9d0e8ea45a78b3a425ac9bf7a6991/atp_players.csv`

[6] `https://github.com/serve-and-volley/atp-world-tour-tennis-data`

1.2 Football

1.2.1 Football dataset

Statsbomb provides an open-source dataset[7], which includes over 2500 matches with event data. Certain tasks are based on specific competitions, an overview of the competitions (including competition IDs) can be found under open-data/data/competitions.json. Match-ID x competition mappings can be found under `open-data/data/matches/`. Apply the Elo Algorithm from Task 1 to the 2015/2016 Bundesliga season.

1.3 Cricket

1.3.1 On Cricket in General

The basic rules of the One-Day-International (ODI) in cricket can be found here. We list the key rules that will be the most useful context for the assessment below. For an ODI:

1. Each team plays one inning consisting of 50 overs, with 6 deliveries per over. The team who bats first is determined by a coin toss.

2. A 'win' is recorded when one side scores more runs than the opposing side and all the innings of the team that has fewer runs have been completed. The side scoring more runs has 'won' the game, and the side scoring fewer has 'lost'. If the match ends without all the innings being completed, the result may be a tie or no result.

3. There is theoretically no limit to the number of runs that can be earned with a single hit as the run tally can increase as the striker and non-striker run to opposite ends of the pitch. However, a hit that bounces and reaches the boundary is an automatic 4 runs and a hit that hits or passes the boundary without a bounce is an automatic 6 runs.

4. A 'wicket' is cricket's equivalent to an out in baseball. A batter continues batting until they are out. The main ways a wicket is earned are by a batter being dismissed by the bowler (e.g., bowled out, leg before wicket, etc.), fielded out by a caught ball, or thrown out during a run on the pitch. Note that either runner (not only the batter) can be thrown out and result in an earned wicket for the other team.

5. Each team has 10 wickets, and once all the wickets are lost, the inning ends, whether the 50 overs have been completed or not.

If you still feel like you need more grounding in the game of cricket, you can take 17 minutes of the assessment time to watch Netflix's *Explained: Cricket*, which can be watched for free on YouTube at this link.

[7] Github: `https://github.com/statsbomb/open-data/tree/master`

1.3.2 Cricket dataset

To get started, download the One Day International match results (link here) and ball-by-ball innings data (link here). These data were sourced from cricsheet.org and include ball-by-ball summaries of ODIs from 2006 to the present for men, and 2009 to the present for women. Both data sets are in JSON format, which have been compiled from the source YAML files on cricsheet. A full description of the source data structure and definition of variables is available here.

1.3.3 Cricket-Specific Tasks

Task C1: Building descriptive cricket summaries

1. Determine the win records (percentage win and total wins) for each team by year and gender, excluding ties, matches with no result, or matches decided by the DLS method in the event that, for whatever reason, the planned innings can't be completed.

2. Which male and female teams had the highest win percentages in 2019? Which had the highest total wins? Were these teams the same as those with the highest win percentages? Comment on why the leaders of these two stats might differ.

Task C2: Visualisation of key-characteristics in cricket Setting aside individual batter production, cricket teams have two main 'resources' for producing runs: overs and wickets. The role resources have on run production is central to the statistical method known as 'DLS', which is used to award winners in the case of incomplete/disrupted matches.

1. Use the ball-by-ball summaries under the innings descriptions of each men's match to make a dataset with the run and wicket outcomes for each delivery in a match, excluding matches with no result.

2. Develop a single visualization to show how the total run rate per over (total for all bowls in an over) varies with inning order, remaining overs (out of 50), and remaining wickets (out of 10) at the start of the over.

3. Summarize your conclusions.

Task C3: Expected runs per over model

1. Based on your observations in Task 2, develop and evaluate a model to predict an average team's expected runs per over.

2 Expected Goals Model in Football

Expected Goals is an established and well-researched [2, 7] concept in football (as well as in other sports).[8] .Although major limitations (e.g. [4]) have been raised, xG has a significant predictive power for future success.[9]

2.1 Task 1: Expected Goals (xG) Model in Football

Use the above open-source data to train and evaluate an xG model. A basic model should be built using event data only, however, additional features could be built based on 360 data. Compare different machine learning models (e.g. a logistic regression, xGboost, ...), different features as well as different strategies to avoid overfitting against appropriate validation metrics (e.g. AUC, Ranked Probability Score, ...).[10]

2.2 Task 2: Expected Points / Season Simulations using xG

Expected goals can be used to quantify player and team performance. Assuming an average goalkeeper performance as well as an average finishing performance (i.e. conversion rate xG into goals), xG can help to quantify the factor of luck for football matches: Simulate football matches from a chosen season of the above data-set by modeling each shot as a Bernoulli-Experiment (using the respective xG value as a weight). Summarize the outcome probabilities of each match to a season ranking, i.e. for each team a likelihood of finishing the season at a certain position.

Note that similar concepts are often referred to as expected points.

3 Submission Guidelines

Since the purpose of this project is to showcase your technical and problem-solving skills, please include clear, efficient, and well-organized code along with explanations on the justification for your problem-solving approach, its limitations, and the conclusions you're able to draw at each step.

- Prepare a well-documented **Jupyter Notebook** or **RMarkdown report**, including your code, visualisations and explanations, i.e. both the source file as well as a PDF or HTML version of the notebook showing all cells evaluated.
- You may use any online resources or additional software (with citations) in your work.

[8] A high-level explanation can be found here: `https://www.youtube.com/watch?v=_oL0q62fc_s&themeRefresh=1`

[9] More details: `https://www.americansocceranalysis.com/home/2022/7/19/the-replication-project-is-xg-the-best-predictor-of-future-results`)

[10] Note that xG is an established concept in various sports (e.g. also in Handball [1])

- Clearly label each section of the notebook/report corresponding to the tasks above.
- Make sure to provide detailed explanations of your findings, insights, and any decisions made during data preprocessing and modeling.

References

[1] Michael Adams et al. "Expected Goals Prediction in Professional Handball using Synchronized Event and Positional Data". In: *MMSports 2023 - Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports, Co-located with: MM 2023* (2023), pp. 83–91. DOI: 10.1145/3606038.3616152 (cit. on p. 6).

[2] Gabriel Anzer and Pascal Bauer. "A Goal Scoring Probability Model based on Synchronized Positional and Event Data". In: *Frontiers in Sports and Active Learning (Special Issue: Using Artificial Intelligence to Enhance Sport Performance)* 3.0 (2021), pp. 1–18. DOI: 10.3389/fspor.2021.624475. URL: https://www.frontiersin.org/articles/10.3389/fspor.2021.624475/full (cit. on p. 6).

[3] Arthur Berg. "Statistical Analysis of the Elo Rating System in Chess". In: *CHANCE* 33.3 (July 2020), pp. 31–38. ISSN: 0933-2480. DOI: 10.1080/09332480.2020.1820249. URL: https://www.tandfonline.com/doi/abs/10.1080/09332480.2020.1820249 (cit. on p. 1).

[4] Jesse Davis, Pieter Robberechts, and K U Leuven. *Biases in Expected Goals Models Confound Finishing Ability*. Tech. rep. URL: https://www.espn.com/soccer/story/_/id/37577474 (cit. on p. 6).

[5] Roberto Gásquez and Vicente Royuela. "The Determinants of International Football Success: A Panel Data Analysis of the Elo Rating". In: *Social Science Quarterly* 97.2 (June 2016), pp. 125–141. ISSN: 15406237. DOI: 10.1111/ssqu.12262 (cit. on p. 1).

[6] Mark E Glickman. "Example of the Glicko-2 system". In: *Boston University* 28 (2012) (cit. on pp. 1, 2).

[7] James H. Hewitt and Oktay Karakuş. "A Machine Learning Approach for Player and Position Adjusted Expected Goals in Football (Soccer)". In: (Jan. 2023). DOI: 10.1016/j.fraope.2023.100034. URL: http://arxiv.org/abs/2301.13052http://dx.doi.org/10.1016/j.fraope.2023.100034 (cit. on p. 6).

[8] Im SangHyuk and Chang-Hoon Lee. "World Tennis Number: The new gold standard, or a failure?" In: *Coaching & Sport Science Review, International Tennis Federation* 91 (2023) (cit. on p. 1).

[9] David J. Marcus. "New Table-Tennis Rating System". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 50.2 (July 2001), pp. 191–208. ISSN: 1467-9884. DOI: 10.1111/1467-9884.00271. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/1467-9884.00271https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9884.00271https://rss.onlinelibrary.wiley.com/doi/10.1111/1467-9884.00271 (cit. on p. 1).