Bundesliga Pythagorean Prediction*

Estimating league-specific coefficients (2010/11–2023/24) and prospective PiT accuracy in 2024/25

John Zhang

October 18, 2025

We study a four-parameter Pythagorean points model for the German Bundesliga. Match-level CSVs are cleaned and standardized, season tables are constructed, and end-of-season (EoS) team totals are pooled across 2010/11–2023/24. We estimate league-specific coefficients (a, b, c, d) by minimizing mean absolute error (MAE) using a multi-start Nelder–Mead procedure. Generalization is evaluated with leave-one-season-out (LOSO) validation, summarizing accuracy by the median MAE and Pearson correlation between predicted and realized EoS points. Using the fitted coefficients, we generate "points-in-table" (PiT) forecasts for 2024/25 at 9, 18, and 27 rounds by combining realized points to date with Pythagorean-based projections for remaining fixtures. We report accuracy at each checkpoint, visualize predicted vs. actual points, and highlight the largest over- and under-performers.

Table of contents

*Code and data: https://github.com/Clearsky21z/Bundesliga_Pythagorean_Prediction

1 Introduction

A simple, durable idea in sports analytics is that scoring and conceding contain most of the information needed to forecast a team's points total. In football this is commonly operationalised with a "Pythagorean" relationship between goals for (GF), goals against (GA), and expected points. Following the presentation in Chapter 5 of the Soccer Analytics textbook and Professor Clive Beggs' work applying Pythagorean points to league football (Beggs (2024)), we tailor a four-parameter variant to the German Bundesliga and evaluate how well it predicts the final table during the 2024/25 season.

This paper makes three contributions. First, we construct a reproducible pipeline from raw match CSVs to end-of-season (EoS) team tables for the modern Bundesliga era (2010/11–2023/24). Second, we estimate **league-specific** coefficients ((a,b,c,d)) by minimizing mean absolute error on a pooled team–season panel, and we quantify transportability with leave-one-season-out (LOSO) validation. Third, we assess **prospective** "points-in-table" (PiT) forecasts for 2024/25 after 9, 18, and 27 rounds, combining realized points to date with Pythagorean projections for the remaining fixtures.

Methodologically we follow the four-parameter form popularized in the football literature: the expected points for team i are $\widehat{\text{PTS}}_i = a \cdot \text{frac}_i \cdot \text{PLD}_i$ where $\text{frac}_i = \dfrac{GF_i^{\,b}}{GF_i^{\,c} + GA_i^{\,d}}$.

The paper proceeds as follows. Section Section 2 formalizes the model and the reproducible pipeline used to clean raw CSVs, construct season tables, estimate coefficients, and perform validation. Section Section 3 reports fitted coefficients, in-sample diagnostics (Table 2), LOSO validation (Table Table 3), and prospective PiT performance for 2024/25 (Table Table 4; Figures Figure 1–Figure 3). Section Section 4 interprets the findings, limitations, and extensions. All artifacts are read from the `output/` directory.

2 Methodology

2.1 Model

For team–season observation i, with goals for GF_i, goals against GA_i, and matches played PLD_i, we use

$$\text{frac}_i = \frac{GF_i^{\,b}}{GF_i^{\,c} + GA_i^{\,d}}, \qquad \widehat{PTS}_i = a \cdot \text{frac}_i \cdot PLD_i.$$

Here $a > 0$ scales the expected fraction to the 3-points-per-win system, and $b, c, d > 0$ control curvature and the relative impact of scoring vs conceding.

Estimation. Let y_i be realized EoS points and $\widehat{y}_i(a, b, c, d)$ the model prediction. We choose (a, b, c, d) to minimize mean absolute error (MAE) across the pooled team–season set (2010/11–2023/24) using **Nelder–Mead** with multiple starts.

Validation. We perform **leave-one-season-out (LOSO)** validation: for each season s, refit on all other seasons and evaluate on s. We summarize performance by median MAE and median Pearson correlation r across held-out seasons.

2.2 Data

2.2.1 Data Overview

This study uses match-level Bundesliga CSV files covering the modern era (2010/11–2024/25). Each file contains all fixtures for a season with home/away teams, goals scored, and the full-time result. We transform these match records into season-level team summaries (end-of-season, EoS) and also build a stacked panel of team–season observations for 2010/11–2023/24 to estimate league-specific Pythagorean coefficients. The 2024/25 season is reserved for *prospective* evaluation using points-in-table (PiT) forecasts after 9, 18, and 27 rounds.

2.2.2 Variables and Measurements

1. **Season**: File-level season tag (e.g., `D1_2010_2011.csv`) used as the season identifier.
2. **Team**: Club name in that season (character string, normalized in cleaning).
3. **PLD**: Matches played by the team in the league season (integer; Bundesliga: 34).
4. **GF**: Goals For — total goals scored by the team across league matches (integer).
5. **GA**: Goals Against — total goals conceded by the team across league matches (integer).
6. **PTS**: League points under the 3–1–0 system (3 for a win, 1 for a draw, 0 for a loss), computed from match results (integer).

These variables are sufficient to compute the Pythagorean fraction and the implied points expectation without any external ratings or betting information.

2.2.3 Dataset Selection

We estimate coefficients on the pooled panel of team–seasons from **2010/11–2023/24** only. Those years provide a consistent competitive format (18 teams, 34 rounds) and align with the time span used in our scripts. All seasons with clean, complete data in that window are included. The **2024/25** season is held out entirely for prospective testing: we form PiT predictions after 9, 18, and 27 rounds and compare them with the final 2024/25 table.

2.2.4 Data Processing Tools

All steps are scripted in **R**:

- `script/01-data_cleaning.R`: drop empty columns, normalize team names, parse dates/times, and recompute `FTR` from goals.
- `script/02-data_aggregating.R`: build per-season EoS tables and a pooled stack.
- `script/03-fitting_coefficients.R`: filter to 2010/11–2023/24, fit ((a,b,c,d)) by MAE (multi-start Nelder–Mead), compute in-sample and LOSO metrics; writes CSVs to `output/`.
- `script/04-pit_testing_2024_25.R`: generate PiT evaluations for 2024/25 (Rounds 9/18/27) and save plots/tables to `output/`.

Core packages: Wickham et al. (2023), Wickham, Hester, and Bryan (2024), Wickham (2016), Müller (2020), Hester, Wickham, and Csárdi (2024), Wickham (2023), Xie (2014).

2.2.5 Example: Bundesliga 2010/11 End-of-season Table

Table 1 shows the final Bundesliga standings for 2010/11 season as reconstructed from the cleaned match file.

Table 1: Bundesliga 2010/11 End-of-season Table

Team	PLD	GF	GA	GD	PTS
Dortmund	34	67	22	45	75
Leverkusen	34	64	44	20	68
Bayern Munich	34	81	40	41	65
Hannover	34	49	45	4	60
Mainz	34	52	39	13	58
Nurnberg	34	47	45	2	47
Kaiserslautern	34	48	51	-3	46
Hamburg	34	46	52	-6	45
Freiburg	34	41	50	-9	44
FC Koln	34	47	62	-15	44
Hoffenheim	34	50	50	0	43
Stuttgart	34	60	59	1	42
Werder Bremen	34	47	61	-14	41
Schalke 04	34	38	44	-6	40
Wolfsburg	34	43	48	-5	38
M'gladbach	34	48	65	-17	36
Ein Frankfurt	34	31	49	-18	34
St Pauli	34	35	68	-33	29

3 Results

3.1 Fitted coefficients (2010/11–2023/24)

The pooled Bundesliga coefficients are **a=2.4177, b=1.2318, c=1.1785, d=1.2174** (see `output/bundesliga_coefs_pooled.csv`).

3.1.1 In-sample diagnostics

- **MAE**: 2.98
- **Correlation (r)**: 0.968

Table 2: In-sample metrics by season (2010/11–2023/24)

Season	MAE	r
D1_2010_2011.csv	3.67	0.931
D1_2011_2012.csv	2.98	0.976
D1_2012_2013.csv	2.94	0.971
D1_2013_2014.csv	3.58	0.969
D1_2014_2015.csv	2.89	0.973
D1_2015_2016.csv	1.64	0.992
D1_2016_2017.csv	3.66	0.950
D1_2017_2018.csv	2.95	0.971
D1_2018_2019.csv	2.80	0.975
D1_2019_2020.csv	2.62	0.977
D1_2020_2021.csv	2.84	0.973
D1_2021_2022.csv	2.94	0.966
D1_2022_2023.csv	3.62	0.952
D1_2023_2024.csv	2.61	0.986

3.1.2 Leave-one-season-out (LOSO)

Overall across held-out seasons:

- **Median MAE**: 3.00
- **Median r**: 0.972

Table 3: LOSO per-season holdout performance

Season	a	b	c	d	MAE	r
D1_2010_2011.csv	2.401115	1.227495	1.170509	1.212918	3.70	0.931
D1_2011_2012.csv	2.490733	1.232563	1.186197	1.226450	3.00	0.975
D1_2012_2013.csv	2.133682	1.162348	1.071932	1.120948	3.20	0.969
D1_2013_2014.csv	2.225726	1.201309	1.126899	1.165948	3.72	0.968
D1_2014_2015.csv	2.751209	1.240201	1.216807	1.260475	3.07	0.973
D1_2015_2016.csv	2.286555	1.226001	1.157699	1.198718	1.69	0.992
D1_2016_2017.csv	2.403808	1.227016	1.170236	1.212843	3.70	0.951
D1_2017_2018.csv	2.402950	1.233615	1.178997	1.217211	2.95	0.971
D1_2018_2019.csv	2.471014	1.226372	1.177819	1.218519	2.82	0.975
D1_2019_2020.csv	2.175335	1.230578	1.151851	1.189216	2.68	0.976
D1_2020_2021.csv	2.492685	1.274063	1.230646	1.263104	2.99	0.972
D1_2021_2022.csv	2.487198	1.222908	1.175689	1.217528	2.99	0.966
D1_2022_2023.csv	2.296429	1.254117	1.190680	1.222013	3.70	0.951
D1_2023_2024.csv	2.403005	1.233461	1.178837	1.217079	2.62	0.986

3.2 Prospective 2024/25 PiT

PiT projects the final table using realized points to date plus Pythagorean projections for remaining fixtures under the fitted coefficients.

Table 4: Bundesliga 2024/25 PiT summary after 9, 18, and 27 rounds.

Rounds	r	MAE
9	0.833	6.91
18	0.942	4.13
27	0.960	3.32

3.2.1 Visual comparisons (predicted vs actual)

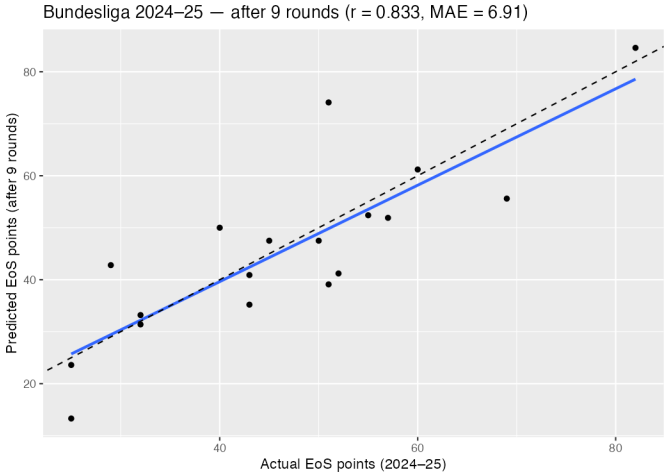

Figure 1: Predicted vs actual EoS points using PiT after Round 9 (2024/25).

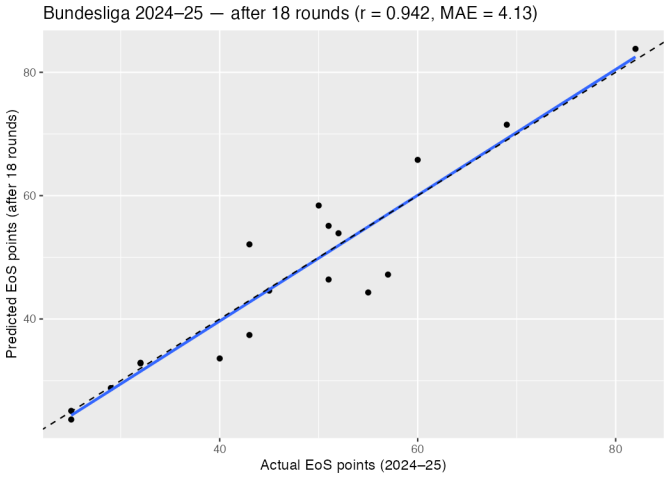

Figure 2: Predicted vs actual EoS points using PiT after Round 18 (2024/25).

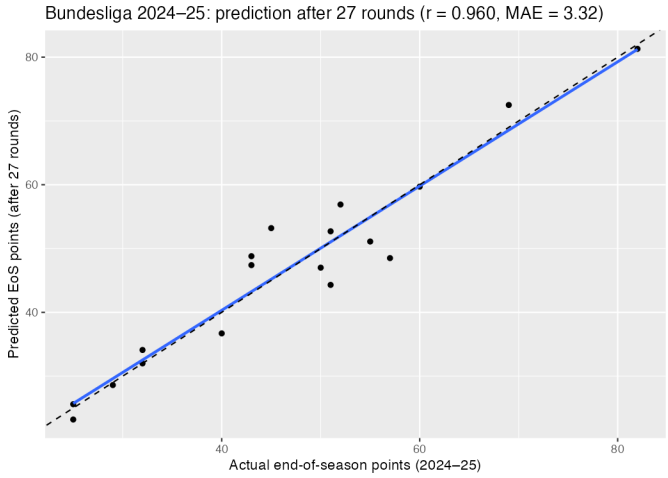

Figure 3: Predicted vs actual EoS points using PiT after Round 27 (2024/25).

3.2.2 Biggest over/under at Round 27

Table 5: Top-5 absolute deviations between predicted total (after Round 27) and final EoS points, 2024/25.

Team	Predicted_Total	Actual_PTS	Error	Abs_Error	Direction
Dortmund	48.5	57	-8.5	8.5	Under (pred<actual)
M'gladbach	53.2	45	8.2	8.2	Over (pred>actual)
Werder Bremen	44.3	51	-6.7	6.7	Under (pred<actual)
Wolfsburg	48.8	43	5.8	5.8	Over (pred>actual)
Mainz	56.9	52	4.9	4.9	Over (pred>actual)

4 Discussion

4.1 Contributions and Findings

This paper calibrates a four-parameter Pythagorean points model to the Bundesliga, following the formulation discussed in Chapter 5 of the Soccer Analytics text and the approach used by Professor Clive Beggs. We estimate league-specific coefficients on seasons 2010/11–2023/24 and evaluate prospective "points-in-table" (PiT) forecasts for 2024/25 at three checkpoints. The historical fit shows that a simple transformation of goals for and goals against explains end-of-season points with small average error. On the hold-out season, forecasts improve as the campaign progresses: early estimates (Round 9) are noisier, while by Round 27 the predictions are close to the realized totals (see Table 4) and Figures Figure 1–Figure 3)). This confirms

a practical message from the literature: across a full league schedule, goal production and prevention are strong summary measures of team performance.

4.2 Limitations

The model projects future points using only current goals for/against and matches played. It does not account for opponent strength, injuries, tactical changes, remaining home/away mix, red cards, or schedule congestion. As a result, it will rate a traditionally strong club poorly during an extended slump until the goal balance turns, and it will continue to credit an over-performing side until their goals regress. These omissions are most visible early in a season when fixture lists can be uneven.

4.3 Future Directions

Several extensions would make the model more realistic. First, generate opponent-aware projections for remaining fixtures using xG or rating models and simulate the schedule so home/away and strength of schedule are reflected. Second, allow team form to evolve over time and let (a,b,c,d) drift modestly across seasons via a hierarchical fit. Third, replace pure MAE with a robust loss and report uncertainty—bootstrap or Bayesian intervals—so PiT outputs include ranges, not just points. Finally, in the first 5–10 rounds, shrink expectations toward league averages or pre-season ratings to avoid overreacting to short runs.

References

Beggs, Clive. 2024. *Soccer Analytics: An Introduction Using r*. CRC Press.

Hester, Jim, Hadley Wickham, and Gábor Csárdi. 2024. *Fs: Cross-Platform File System Operations Based on 'Libuv'*. https://CRAN.R-project.org/package=fs.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2023. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.