DATA-DRIVEN PERFORMANCE ANALYSIS OF CHELSEA FC

Group 17
Business Data Mining

DEBASMIT BORA
ADITYA NARAYAN
JINGBO WANG
TIANLAN XU

BACKGROUND

Chelsea FC is one of England’s most successful modern-era teams, known for its strong international squad and major financial investment.

Founded in 1905 in the Chelsea borough of London, the club has built a reputation for competitiveness both domestically and in Europe.

Chelsea has won multiple Premier League titles, FA Cups, and two UEFA Champions League trophies (2012 and 2021), establishing itself as a top-tier club.

Despite Chelsea's strong historical achievements, recent performance has shown considerable inconsistency.

Chelsea has heavily invested in rebuilding the squad in recent seasons, yet on-field performance remains unstable.

Despite these investments, results continue to fluctuate, suggesting deeper issues related to player form and opponent strength.

This raises some critical questions:

How do we do better?

Problem Statement

To address this, a data-driven approach is required to:

- **Evaluate the impact of opponent strength on Chelsea's performance**
- **Understand past performances and forecast the future to identify challenges**
- **Quantitatively measure the relationship between team players' form and team results**
- **Identify weak positions or areas needing reinforcement**
- **Use statistical patterns to support transfer strategy decisions**

METHODOLOGY OVERVIEW

Time-Series Forecasting (SARIMA + Fourier Terms)

Used to model and forecast Chelsea's and other PL teams' ELO rating over time, capturing both trend and seasonality in team performance.

Clustering (Player Performance Groups)

Groups players into natural performance types using offensive, defensive, and progressive metrics to identify strong and weak positional clusters.

Weighted Averages (Seasonal Player Stats)

Combines multiple seasons of data with recency weighting to produce stable and realistic estimates of each player's 2025 performance profile.

Linear Regression (Match Result Forecasting)

Predicts match outcomes using numerical features such as ELO, venue, and expected goals to complement time-series and clustering insights.

WHY WE USE SARIMA + FOURIER

Chelsea's ELO rating behaves like a time-series with clear trends and seasonal patterns.
SARIMA captures short-term fluctuations, while Fourier terms model the cyclical performance waves across the Premier League season.
This combination is more accurate and stable than simple moving averages or one-step regressions.

WHAT WE GAIN FROM THIS

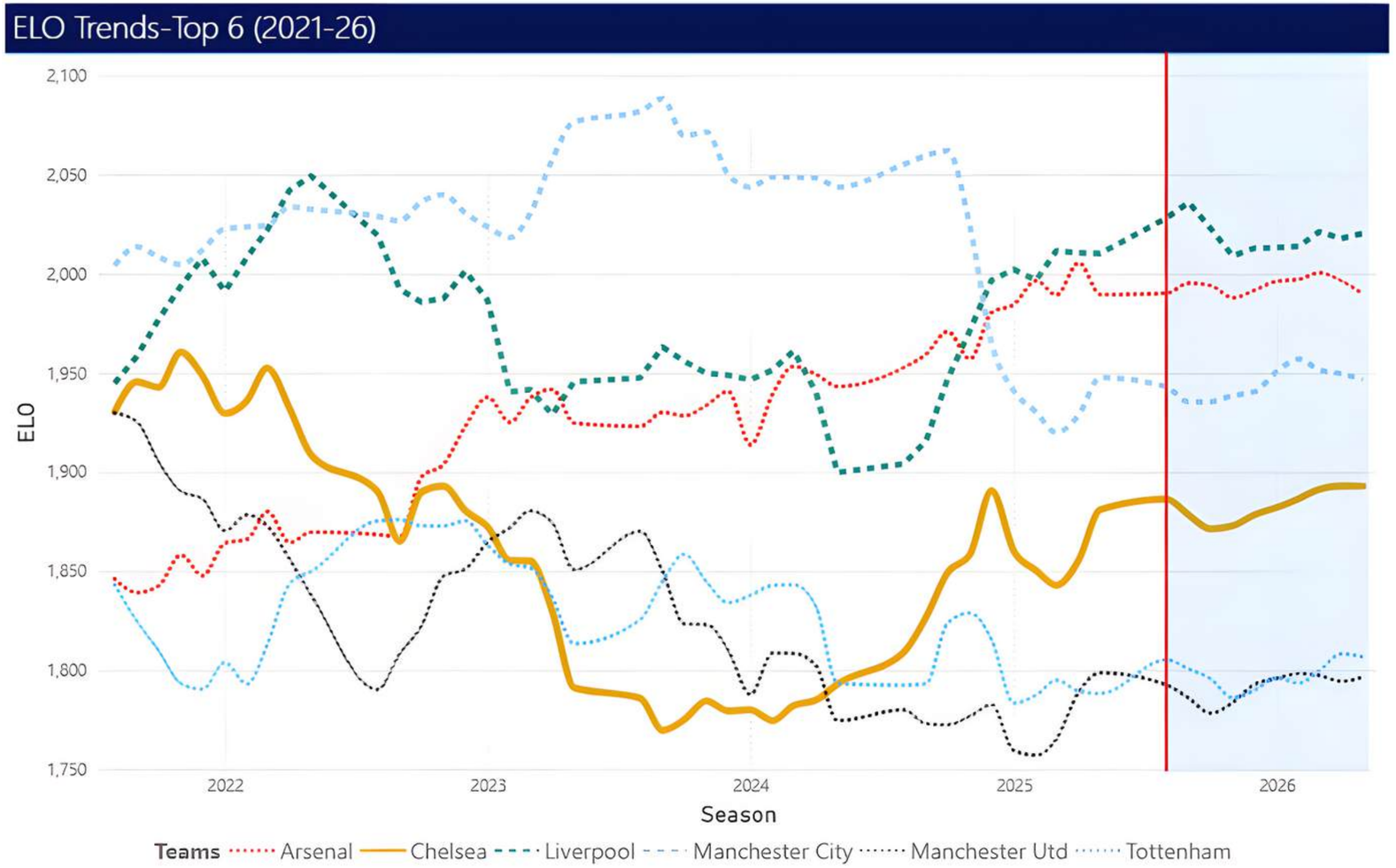
Forecasting ELO allows us to anticipate Chelsea's future strength profile.
We can identify likely peaks and dips across the season, evaluate fixture difficulty more realistically, and detect high-risk periods where performance may drop.
This helps contextualise later predictions and improves match-level forecasting.

SARIMA + FOURIER TERMS (ELO FORECASTING)

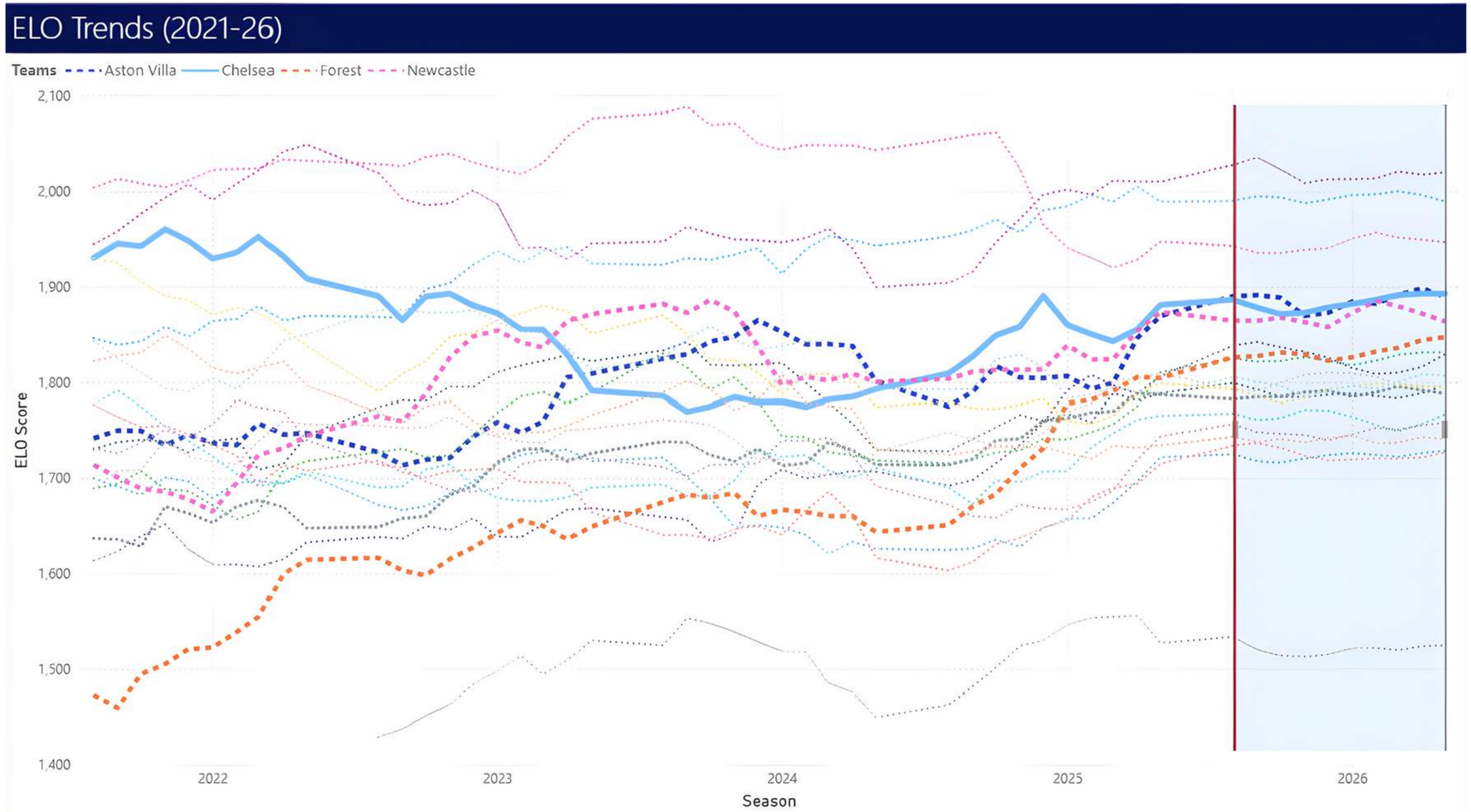
Chelsea vs The Top 5

The EPL Top 5 has traditionally vied for the top spot each season.
From 2015, the Top 5 became the Top 6 as Tottenham became a serious contender.
Over the past few seasons, we have seen fortunes change, including at Chelsea.

To visualize these changing dynamics, we track ELO scores for all PL teams over the past 4 seasons, and use SARIMA to predict the new one.



Chelsea vs The Top 5



Using the same methodology, we can also track new challengers, same as Tottenham in 2015.

We can also see where Chelsea sits on average in the league.

This keeps us ahead of the curve at understanding who our biggest surprises would be in the coming season.

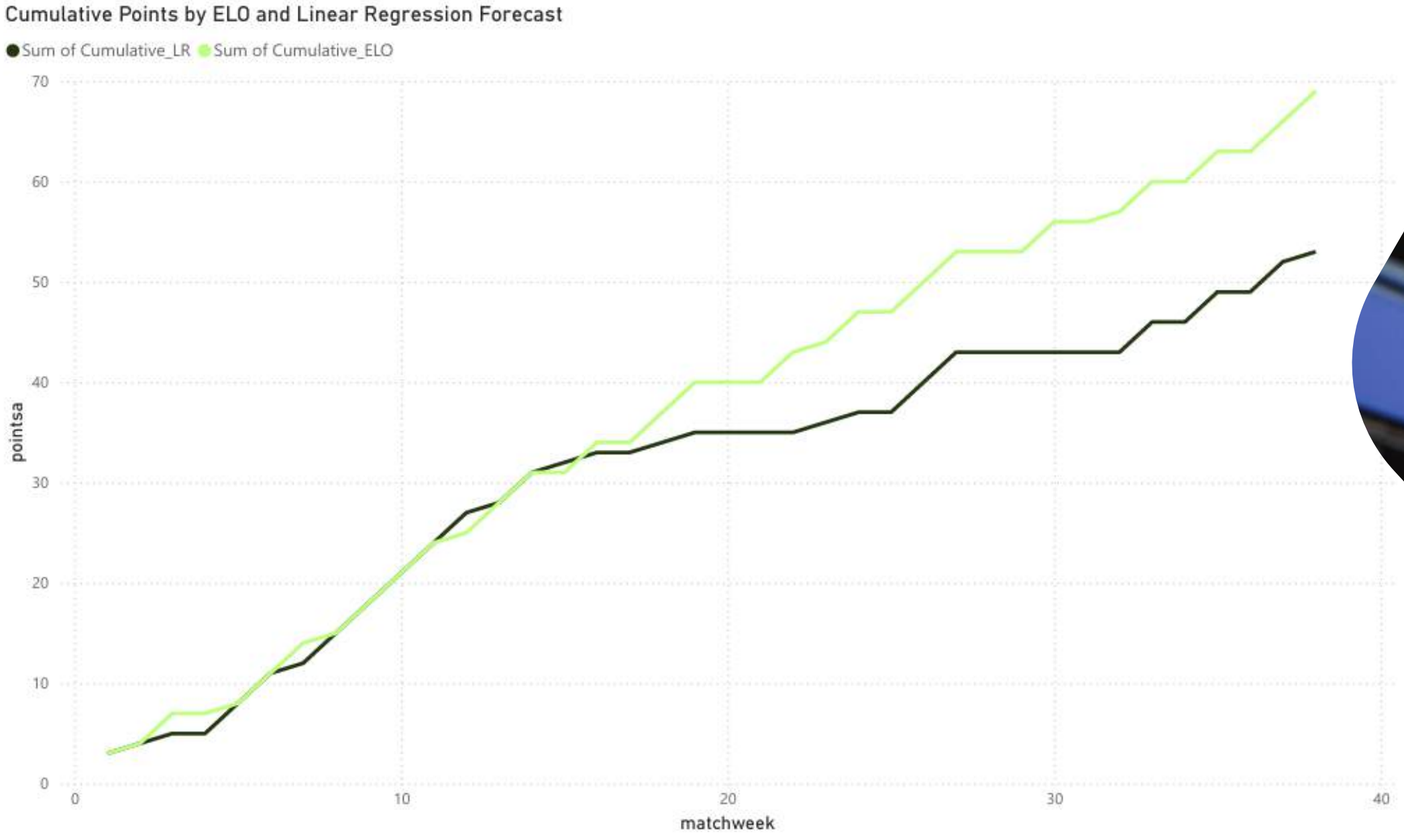
WHY WE USE LINEAR REGRESSION

Match results depend on quantitative features—ELO, home/away effects, expected goals, and opponent strength.
Linear regression captures these relationships cleanly and transparently.
It complements the ELO model by focusing on match-specific numerical drivers rather than long-term team strength.

WHAT WE GAIN FROM THIS

We obtain a second predictive viewpoint to validate or challenge ELO forecasts.
Regression highlights matches where numerical factors predict differently from ELO—often signalling risk, rotation, or inconsistency.
It helps identify fixtures where Chelsea must be more cautious or adjust strategy.

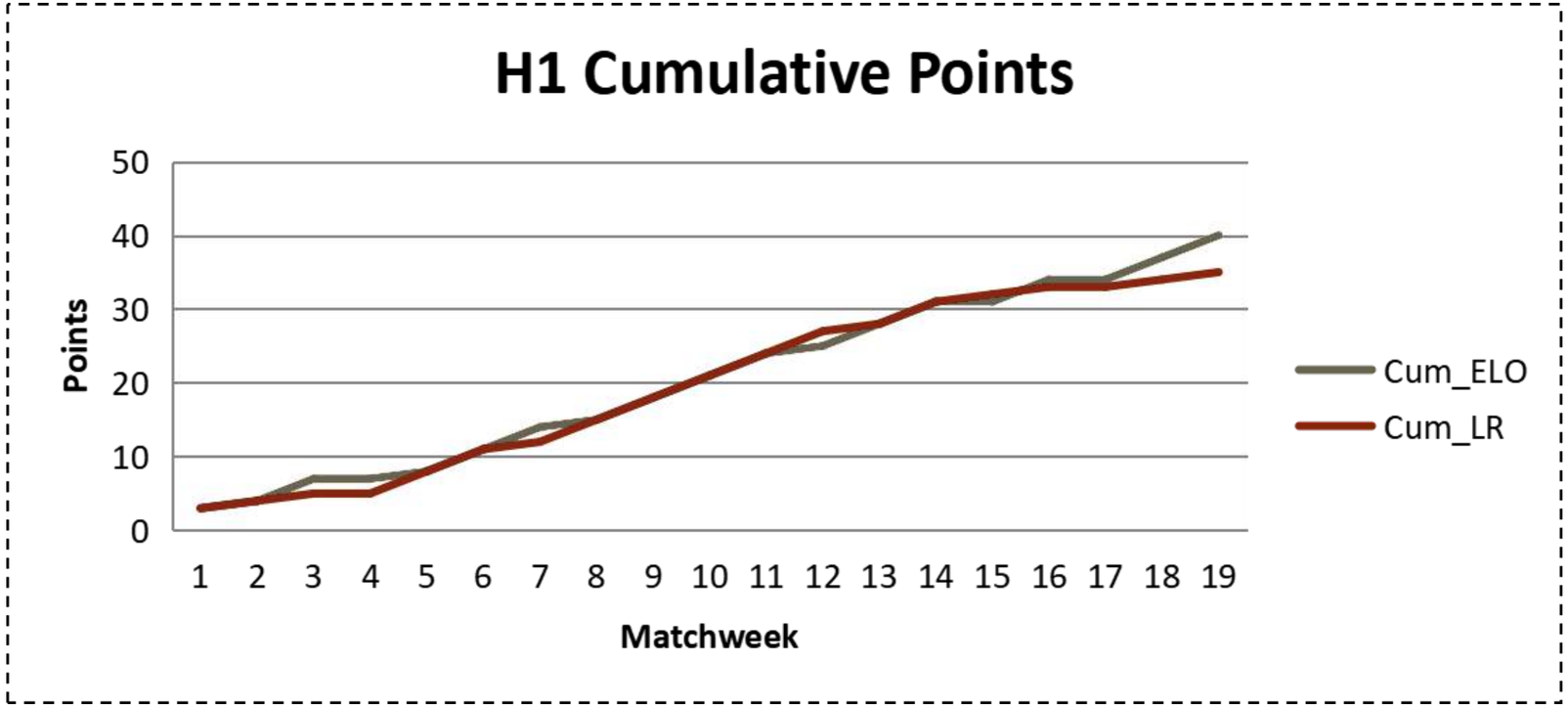
LINEAR REGRESSION (MATCH RESULT FORECASTING)



ELO and LR predicted scores deviate significantly.

ELO forecasts.

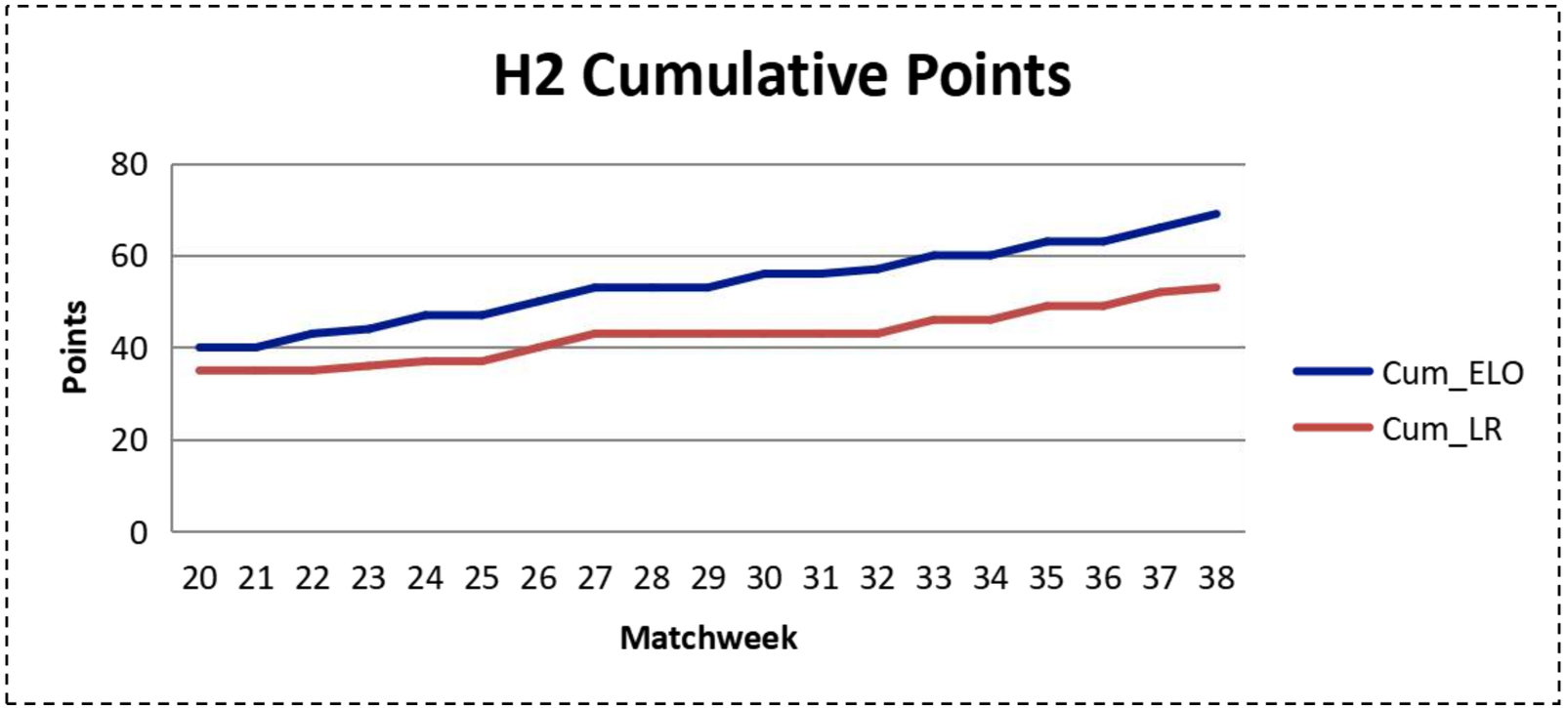
LR uses historical data.



Matchweek	Goals_For	Goals_Against	Result
1	2	1	W
2	1	1	D
3	1	1	W
4	0	1	L
5	2	1	D
6	2	1	W
7	1	1	W
8	2	1	D
9	2	1	W
10	4	0	W
11	2	1	W
12	3	2	D
13	2	2	W
14	2	1	W
15	2	2	L
16	1	1	W
17	1	3	L
18	1	1	W
19	1	1	W

First half of the season predictions stay the same for the first half of the season.

This indicates Chelsea always start strong.
well-rested, eager players play harder and faster.
ELO, a team ability calculation metric, and
LR, a historic trends metric,
agree that Chelsea will start off with good results: 12W, 4D, 3L



Matchweek	Goals_For	Goals_Against	Result
20	0	3	L
21	0	2	L
22	0	1	W
23	2	2	D
24	1	1	W
25	1	3	L
26	1	0	W
27	3	1	W
28	1	3	L
29	1	2	L
30	0	1	W
31	1	2	L
32	1	2	D
33	2	1	W
34	1	2	L
35	3	1	W
36	1	2	L
37	3	0	W
38	1	1	W

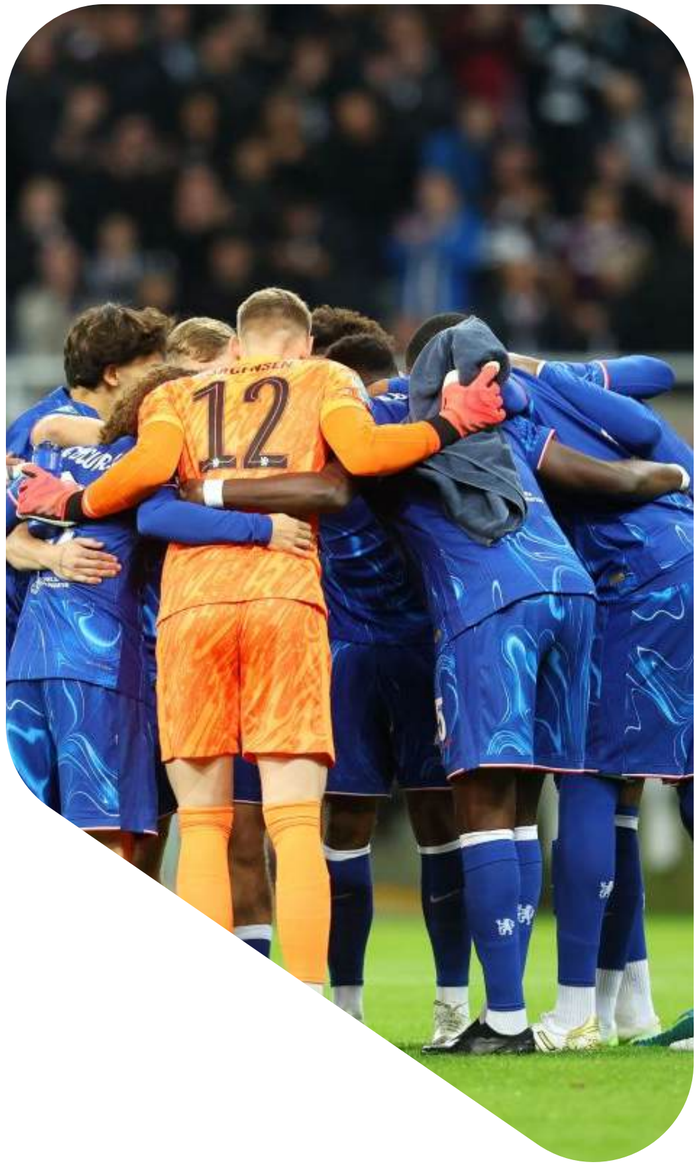
Second half of the season predictions stay the same for the first half of the season.

The deviation is explained by Chelsea's historic record against PL teams, vs their performances in tournaments.

Tourneys contribute to ELO, a good run can greatly boost scores.

LR understands that a good run of form can be marred by a bad VS record, bad signings in the winter, and injuries.

This is where Chelsea need to double down and focus.



POINTS COMPARISON

This KPI section summarizes the total predicted points for Chelsea using two different models: Logistic Regression (LR) and ELO.

Points are calculated based on predicted match outcomes using the standard football scoring rule:

Win = 3 points, Draw = 1 point, Loss = 0 points.

On average, the PL winner is expected to score 80+ points a season.

Chelsea is not predicted to perform very well with the current set of players. Let's see how we can fix that.

53	69
Total LR Points	Total ELO Points

WHY WE USE CLUSTERING

Chelsea players differ widely in offensive output, defensive work, progression, and minutes played.
Clustering groups them into natural performance types instead of relying only on position labels.
This reveals meaningful structures in the squad, highlighting roles that statistics alone cannot explain.

WHAT WE GAIN FROM THIS

We identify which clusters contain key contributors and which represent weak or inconsistent performers.
This exposes structural weaknesses in the squad—forwards lacking output, midfielders lacking balance, or defenders lacking stability.
These insights help inform transfer priorities and tactical adjustments.

CLUSTERING (PLAYER PERFORMANCE GROUPS)

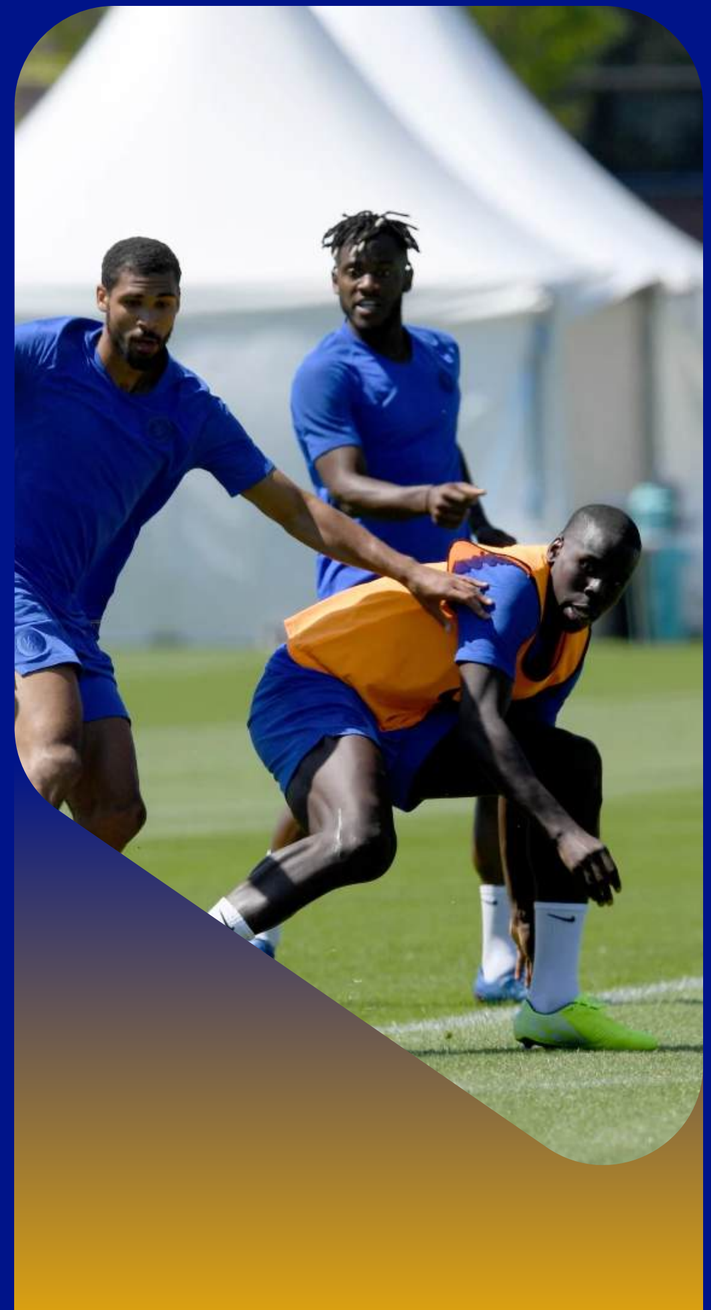
WHY WE USE WEIGHTED AVERAGES

Player performance varies across seasons, and using only last year's stats can misrepresent current ability. Weighted averages give higher importance to recent seasons, smoothing out injuries, outliers, or unusually good/bad years.
This produces a more realistic estimate of each player's current-season level.

WHAT WE GAIN FROM THIS

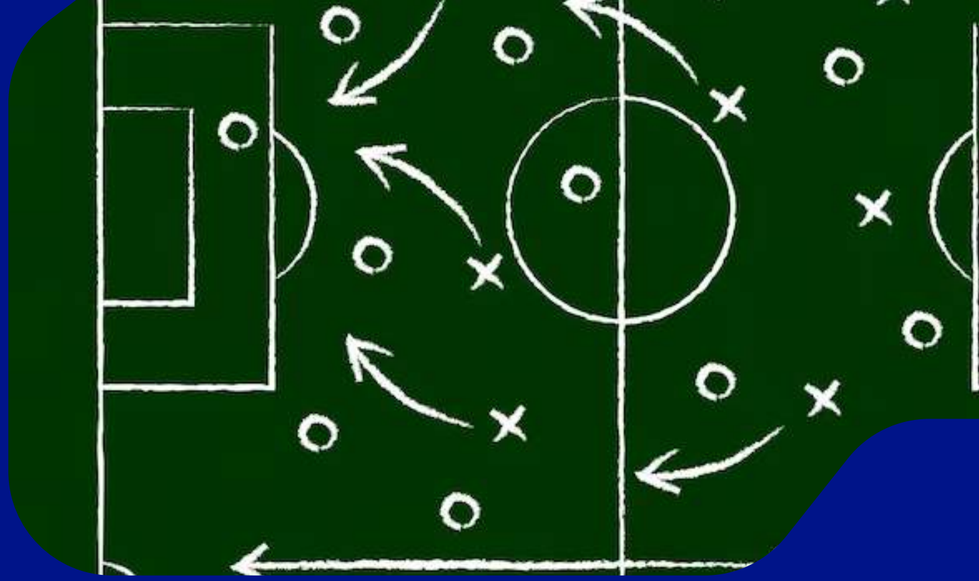
We obtain consistent, up-to-date performance metrics for all Chelsea players in 2025.
These stable stats feed into clustering and regression models, improving prediction quality.
Weighted averages allow us to compare players fairly and identify who is trending up or down.

WEIGHTED AVERAGES (SEASONAL PLAYER STATS)



PLAYING POSITIONS

Roles That Define the Game



Goalkeeper

Prevents goals, commands the box, and often starts counterattacks with throws or kicks.

Defenders

Protect the goal area, intercept passes, and block shots.

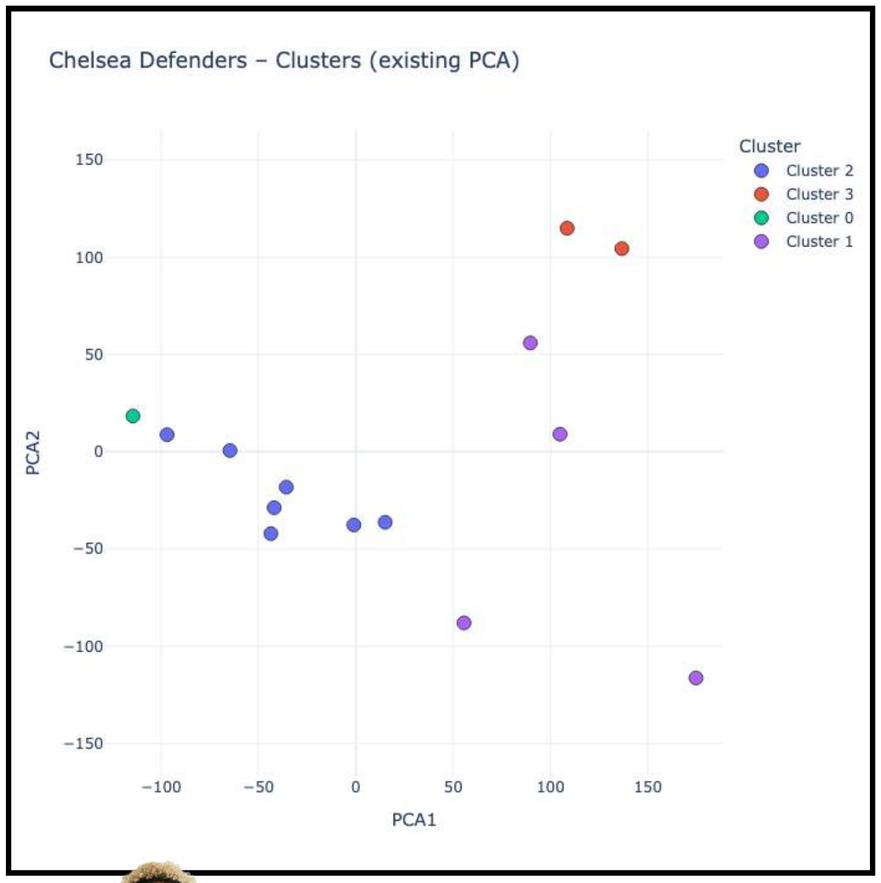
Midfielders

Control play tempo, link defense and attack, and support both offensively and defensively.

Forwards

Primary scorers who lead attacks, create chances, and apply pressure to opposing defenses.

Clustering (Defenders)



3

Reece James:

Tackle Win Rate: 68%

Blocks: 18

Interceptions: 12

Clearances: 26

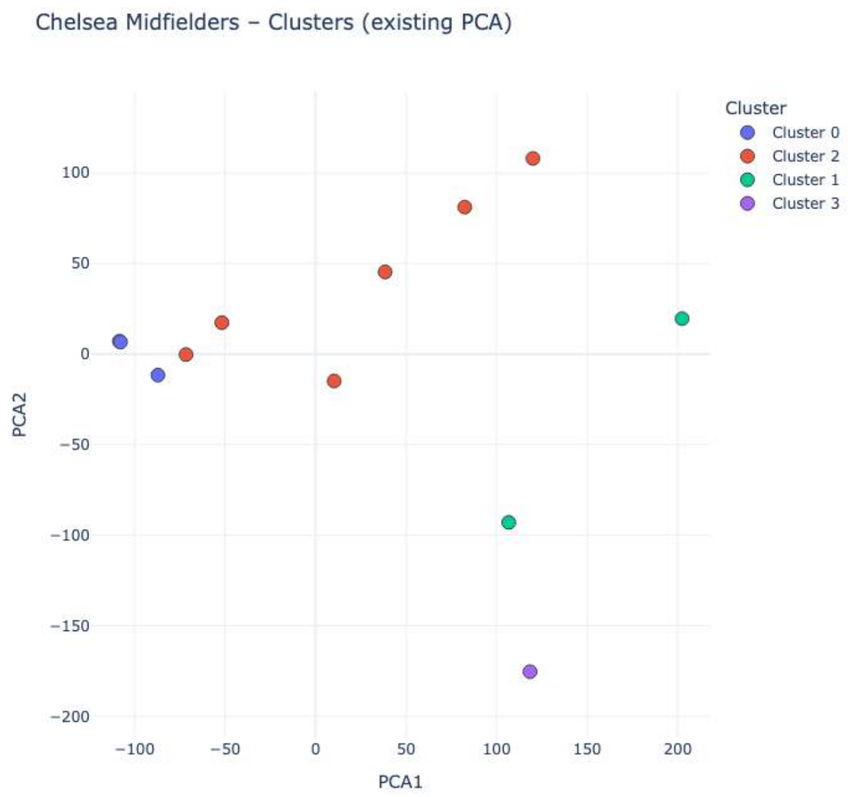
Cluster 0 (Classic CB) : High tackles, Intercepts and Clears. Core Central Defense.

Cluster 1 (Full Backs) : Low Defense, High Progression. Dual Role Players, sharing midfield

Cluster 2 (Rotation Players): Low Matches Played. Needed as Backup

Cluster 3 (Anchors): Key Players, High Stats Around

Clustering (Midfielders)



1

Cole Palmer:

Goals+Assists: 28

Progressive Moves: 324

Tackle Win Rate: 55%

Defensive Contributions: 50

Cluster 0 (Defensive Midfielder): High tackles, Interceptions, Clearances, low goals/assists. Anchors the midfield.

Cluster 1 (Attacking Midfielder): High Assists, G+A, Progression. Playmakers creating chances.

Cluster 2 (Box-to-box Mids): Moderate goals, assists, and defensive stats. Covers lots of ground.

Cluster 3 (Rotation/Utility Mids): Low minutes, mixed stats. Young or backup midfielders.

Clustering (Forwards)

2

Estevao Willian:

Goals: 13

Assists: 9

Progressive Moves: 383

Progressive Passes: 92

Cluster 0 (High goal scorers) : Players with high goals and goals per 90 minutes, fewer assists. Central strikers who finish chances.

Cluster 1 (Playmaking Forwards) : Moderate goals, higher assists and Progressive Carries/Progressive Passes. These might be inside forwards or creative wide forwards.

Cluster 2 (Rotation forwards): Low Matches Played and low stats overall. Often substitutes or young players.

Cluster 3 (Balanced forwards): Moderate goals and assists, some progressive actions. Versatile forwards who contribute in multiple ways.

Analysis of Weak Clusters

Ball-Playing Centre-back

Chelsea need:

- better build-up
- line-breaking passing
- someone to partner Disasi/Colwill reliably

Offensive Center Midfielder

Chelsea have:

- creators (Enzo/Palmer)
- destroyers (Caicedo)

But not a goal-scoring #8 like

- Barella
- Valverde
- Gallagher was the closest, but technically limited.

Elite Finishing Striker

Chelsea still don’t have a:

- pure finisher
- high Gls90 scorer
- reliable xG converter

Jackson is energetic but not clinical.

Broja is gone.

Nkunku is never fit.

This is Chelsea’s biggest gap.

Player Recommendation

1. Alessandro Bastoni - The Elite Ball-Playing Centre-Back

Profile fit: Progressive Passing + Calm Defender

Bastoni gives Chelsea what they currently lack: calmness, control, and progression from the left side.

He fills the *ball-playing CB cluster*, which your analysis identified as a major weakness. His progression metrics (Prg Passes, Prg Carries, long distribution) are elite.

2. Fermín López - The High-Intensity, Goal-Scoring #8

Profile Fit: Cluster A/B hybrid - goal-scoring, high-press interior midfielder

Fermín López is an explosive, high-tempo, attacking midfielder with perfect traits for a modern Premier League interior.

Chelsea badly lack goals from midfield, Fermín fixes that immediately

3. Serhou Guirassy - The Pure Clinical Finisher Chelsea Lack

Profile Fit: High-Output Forward

Chelsea desperately need a striker who *converts* chances. Guirassy is elite at this.

He is a direct upgrade to the Cluster “finisher” archetype that Chelsea do not currently have.

THANK YOU