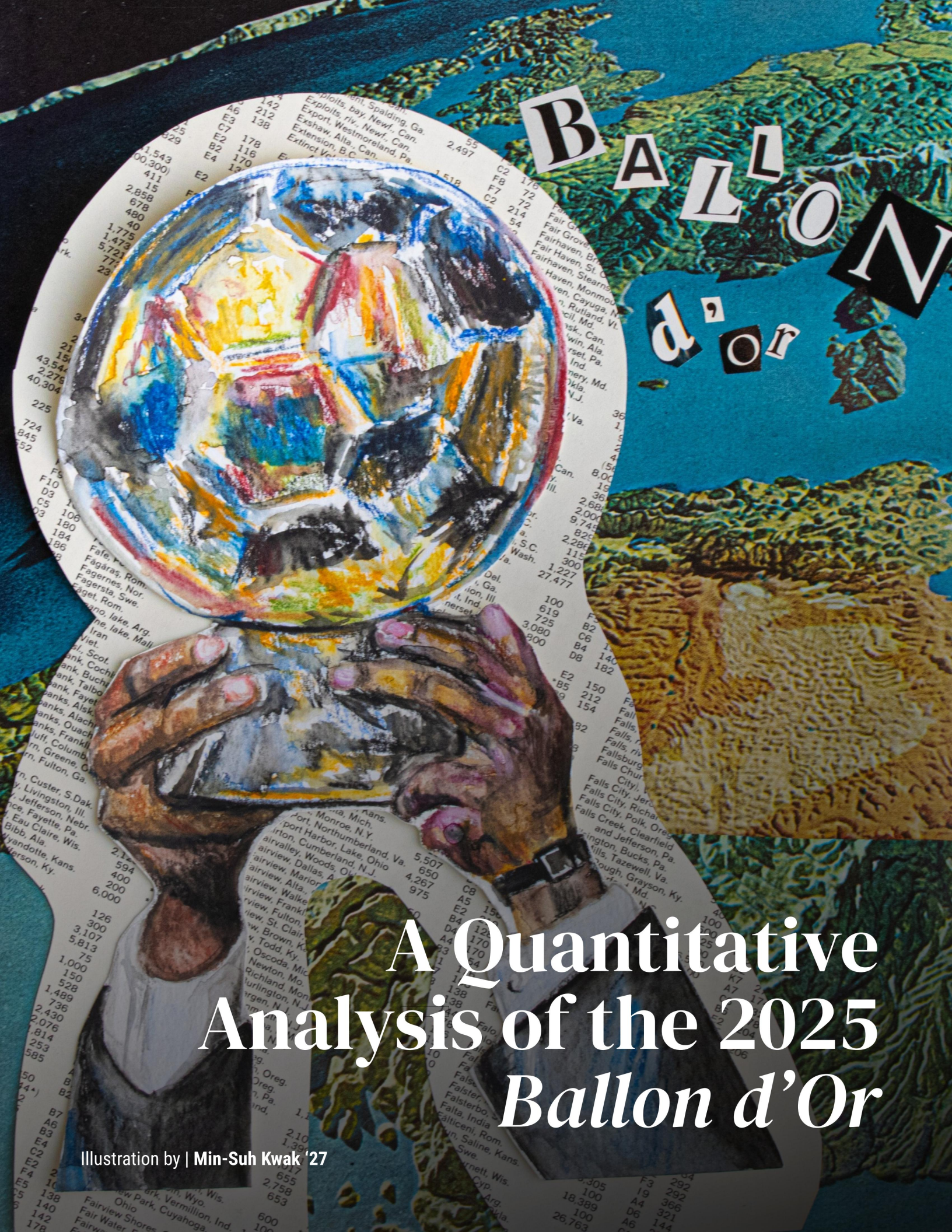


A Quantitative Analysis of the 2025 *Ballon d'Or*

Illustration by | **Min-Suh Kwak '27**

ORIGINAL RESEARCH

A Quantitative Analysis of the 2025 Ballon d'Or

Lucas Olmeta '28 & Ben Kras '28
Washington University in St. Louis

The Ballon d'Or debate often reflects disagreement over whether the award recognizes pure on-pitch performance or a combination of performance, honors, and narrative. This tension implies two distinct analytical goals: evaluating performance-based "deservedness" and modeling how voters actually convert performances and achievements into rankings. We study the 2025 race with a nominee-only dataset and complementary methods—descriptive rankings, historical winner profiling, and within-season vote-share modeling—to address both perspectives. Using ridge regression with softmax normalization to map performance statistics and major trophies to expected voting outcomes, we find that including trophy indicators substantially improves ranking fidelity and identifies the correct winner on the held-out 2024–25 season. Uncertainty analysis via stratified bootstrap reveals clear predicted separation at the top of the ballot with the winner's dominance robust across resampled training sets. Taken together, these approaches show how different defensible framings of "deservedness" yield distinct conclusions and clarify the gap between performance-based expectations and realized voting outcomes.

1. Introduction and Research Questions

1.1 Motivation: Why 2025 is Debated

The Ballon d'Or is intended to recognize the world's best footballer over a season, yet the award's outcome often reflects a blend of individual performance, team success, and narrative. The 2025 race is debated because it plausibly admits multiple "best player" definitions that can yield different winners: (i) best underlying on-pitch contribution independent of trophies, (ii) most decorated season on the biggest stages, or (iii) the player whose season most resembles historically rewarded profiles. For instance, this split is perfectly illustrated by the debate over whether the underlying statistical dominance of a player like Mohamed Salah should outweigh the trophy-laden, high-leverage performances of Ousmane Dembélé. The public disagreement is therefore not merely about disagreements in statistics, but about disagreements in what the award functionally measures and how voters convert performances and achievements into points. This paper treats the 2025 controversy as an empirical question: given the historical voting record and nominee performances, what outcome would be expected under different defensible framings, and how large (or small) are the implied differences between the leading candidates?

1.2 Research Questions

This paper is organized around three research questions designed to separate descriptive performance claims from claims about voter behavior. To achieve this, the study combines descriptive performance analysis, historical winner comparison, and predictive modeling to evaluate the leading candidates.

RQ1 (Deservedness / performance). Among 2024–25 Ballon d'Or nominees, who had the strongest season according to reproducible performance evidence, and how sensitive is that conclusion to reasonable choices about efficiency, volume, and durability?

RQ2 (Closeness / uncertainty). How close was the 2025 race in expected vote share terms, and what is the uncertainty around the predicted separation between the top contenders?

RQ3 (Historical comparability). Was the 2025 winner's season similar to prior winners' when evaluated on consistent, normalized measures?

These questions are answered by triangulating across complementary methods: descriptive rankings, historical winner profiling, nominee-only vote-share modeling, uncertainty quantification, and transparent composite scoring.

1.3 Scope, Definitions, and Nominee-Only Framing

Unit of analysis. The fundamental observational unit is a player-season restricted to Ballon d'Or nominees. Each row corresponds to one nominee's club-season statistical record aligned to the Ballon d'Or voting associated with that season.

Population and selection. This is explicitly a nominee-only study. The dataset does not represent all professional footballers; it represents the subset of players the award ecosystem has already deemed award-relevant. This framing is deliberate: it matches the real decision problem faced by voters (choosing among nominees) and avoids conflating "not nominated" with "not worthy." However, it implies that conclusions should be interpreted as conditional on nomination. Consequently, this restricts the comparison to an already elite tier of footballers, meaning our conclusions evaluate marginal differences at the absolute upper echelon of the sport rather than against the global average.

Positions and exclusions. The main modeling analysis focuses on outfield nominees. Goalkeepers follow different statistical regimes and require a specialized feature set; mixing them into one model would generally degrade interpretability and fairness. Descriptive sections may mention goalkeepers where helpful, but the central comparative claims are for outfield players.

Outcome definition: voting points vs vote share. The award is

decided by voting points, but raw point totals can vary across seasons due to changes in voting scale and electorate. Therefore, the primary modeling target is defined as within-season vote share, i.e., each nominee's fraction of total points awarded that season among the nominee set being analyzed. This choice yields three advantages: comparability across seasons even if point scales change; interpretability as a "probability-like" quantity (nonnegative, sums to one); and natural alignment with the paper's "how close was it?" question, since differences in vote share are directly interpretable as gaps in expected support.

Performance vs résumé variables. We distinguish two conceptually different feature families: (i) performance features derived from match event aggregates (e.g., chance creation, scoring threat, progression) expressed per-90 where appropriate, and (ii) résumé/trophy indicators capturing team achievements and high-salience accolades. This separation supports two legitimate but different conclusions: "best on-pitch season" versus "most consistent with historical voter reward patterns."

1.4 Overview of Parts and Contributions

Part 2 (Data acquisition). We construct a dataset by collecting season-level player statistics from FBref and Ballon d'Or voting outcomes from structured sources.

Part 3 (Descriptive context). We rank nominees on interpretable baseline metrics (e.g., goals/assists, expected goals and assists, totals and per-90 variants) to provide transparent context and establish the main tradeoffs (volume vs efficiency, availability vs peak rate).

Part 4 (Winner profile and similarity). We characterize historically rewarded nominee seasons by examining associations between features and voting outcomes and constructing a past-winner profile. We then measure how similar each current nominee is to the historical winner profile, providing a complementary lens that is descriptive rather than predictive.

Part 5 (Nominee-only vote-share model). We train constrained models on past nominee seasons to predict within-season vote share, producing expected rankings for 2024–25 and decomposing discrepancies between predicted and actual outcomes. We include uncertainty quantification to estimate win probabilities and expected separation between the top contenders, directly addressing "how close was it?"

Part 6 (Synthesis). We discuss all methods to provide a robust conclusion, clearly separating performance-based claims from claims about voter behavior, and we document limitations and future improvements.

Part 7 (References). We list references, mainly our data sources.

1.5 Modeling Assumptions and Constraints

Small-sample regime and regularization. The nominee-only panel spans only a few seasons and a limited number of nominees per season. This imposes a strict bias–variance tradeoff: flexible models risk overfitting idiosyncrasies of a single season. By using regularized models like Ridge regression, we can prevent the algorithm from overreacting to highly correlated football statistics, such as expected goals and shot creating actions, thus reducing the risk of overfitting. Consequently, the modeling strategy prioritizes constrained, regularized approaches and evaluation metrics aligned with the project's decision goal (ranking) rather than only minimizing pointwise error.

Ranking is the primary estimand. The central inferential object is not an exact prediction of raw points, but the implied ordering of nominees and the magnitude of expected vote-share gaps. Therefore, rank-based metrics (e.g., within-season rank correlation, top-*k* inclusion) are treated as primary. Calibration of absolute share values is treated as secondary and is assessed cautiously.

Within-season normalization and comparability. The key comparability device is vote share rather than voting points. This implicitly assumes that, within a season, the nominee vote distribution meaningfully captures relative consensus, and that comparing shares across seasons is more stable than comparing raw totals under potential rule changes.

Season structure and dependence. Nominees within a season are not independent in the sense that their vote shares must sum to one and their narratives and team outcomes are correlated. Uncertainty estimation is therefore designed to respect season structure (e.g., resampling within seasons) rather than treating rows as independent and identically distributed observations.

Feature measurement and role heterogeneity. FBref aggregates are role-dependent (forwards, midfielders, defenders generate different distributions). Without perfect position/role control, any single "best season" ordering risks conflating role with value. The composite score (Part 5) mitigates this by grouping metrics into categories and by enabling position-aware normalization if labels are available, but the nominee-only data still limit fully causal interpretations.

Minutes/durability as both signal and confound. Availability is plausibly part of value (a durable elite season is different from a short peak), but minutes also mechanically inflate totals. The analysis treats durability explicitly rather than letting it leak implicitly through totals alone; however, it remains a core modeling tension and is addressed via per-90 features plus controlled durability terms and ablations.

Omitted-variable bias and narrative factors. Voting incorporates factors not captured by the statistical tables (e.g., international-tournament timing, media narratives, positional expectations, "big moments," leadership, prestige effects). The model should therefore be interpreted as predicting votes conditional on included measurable covariates, not as a complete causal explanation of voting.

2. Data Acquisition

2.1 Data Sources and Seasons

Performance features for all seasons were sourced from FBref via Engelmann, T.'s soccerdata Python package, and joined into one continuous DataFrame[A1.1][A1.2]. While publicly sourced datasets may be limited by their depth, detail, and accuracy, they are also the most accessible and reproducible. Data was collected for the 2021-2022, 2022-2023, 2023-2024, and 2024-2025 seasons. This range was chosen to match the switch in the Ballon d'Or voting format starting in 2021-2022, which switched the voting period from the calendar year to the European club season calendar. This DataFrame was subsequently supplemented through manual data-entry of résumé and trophy features sourced from Transfermarkt.

2.2 Fbref Scraping and Table Coverage

All available performance features were sourced from FBref for both outfield players and goalkeepers. Data ingested includes every player in the FBref database for the aforementioned seasons,

not just those who were nominated for the Ballon d'Or. FBref offers a wide breadth of performance features, including but not limited to, advanced metrics such as expected goals, shot types, pass types, set piece types, defensive actions, and goalkeeping actions.

2.3 Ballon d'Or Voting Points Collection

Ballon d'Or voting point totals for each year were added through manual data-entry. Totals were sourced from Wikipedia [1]. It should be noted that beginning in the 2022-2023 season, the Ballon d'Or voting system was changed from a 5-choice ranking system to a 10-choice ranking system. As a result, the total pool of voting points is larger for the 2022-2023, 2023-2024, and 2024-2025 seasons than it is for the 2021-2022 season.

2.4 Data Schema and Storage (Raw vs Cleaned)

Raw data ingested from FBref was accumulated into a single DataFrame and aggregated by name and season to avoid duplicates. Since observations were split by name and season, each observation is uniquely identifiable through its name and season features. We also cleaned the data by normalizing player names through the removal of accents and most special characters. After this process, we made several different datasets available in the data section of our GitHub repository. Depending on the dataset, they differ by both features and observations. Several of the files are dedicated to specific subsets of observations, such as nominees, outfield nominees, and goalkeeper nominees. These files only include the features relevant to their population, with the outfield nominees file including only outfield features, and the goalkeeper nominees file including only goalkeeping features.

3. Basic Nominee Rankings and Descriptive Context

3.1 Nominee Set and Exclusions

We began by comparing all the 2024-2025 outfield nominees directly based on various basic metrics. This not only served to help gain a better understanding of the data, but also to set a baseline for the rest of the analysis. Both total and rate metrics are included in the descriptive analysis. Calculating totals throughout a season-long time frame allows us to quantify total impact throughout the entire voting window, and rewards players who stay healthy, play consistently throughout the season, and log substantial minutes. On the other hand, rate and per-90 metrics allow us to measure productivity more directly by standardizing output to a common time basis (such as 90 minutes). This helps to control for differences in minutes played, which can be driven by factors outside of a player's underlying quality or resilience (such as certain contact injuries, extra cup games due to team quality, and differing amounts of league games).

3.2 Volume Metrics: Totals

First, these nominees were ranked by several common key metrics accumulated across the entire season, beginning with non-penalty expected goal contributions. Non-penalty expected goal contributions is a common way to quantify the quantity and quality of chances a player has created for themselves and others. It is calculated by summing the probability of a goal resulting from every shot and pass a player takes throughout a season. According to BBC, expected goals "has become commonplace across football analytics" and is "is seen as a good indicator of how a player or team has been performing in front of goal" [2].

3.3 Rate Metrics: Success Percentages and Per 90 Efficiency

To gauge offensive efficiency, metrics such as non-penalty expected goal contributions per 90 and shot creating actions per 90 were plotted and compared on a peer-to-peer basis. These metrics allow us to quantitatively measure offensive efficiency, rewarding players who create chances for both others and themselves. As a measure of efficiency less skewed towards attackers, successful take-on percentage was also calculated and plotted using the following formula:

$$\text{Successful take-on \%} = 100 \times \frac{\text{Successful take-ons}}{\text{Take-ons attempted}}.$$

3.4 Visualization Summaries for 2024-25

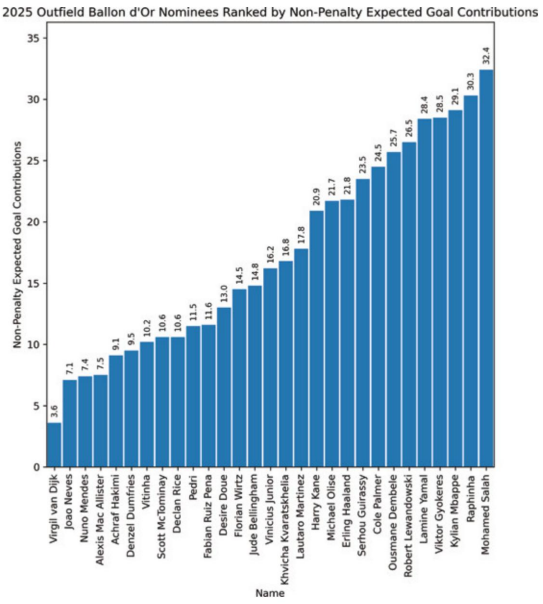


Fig. 1 | 2024-2025 outfield nominees ranked by expected non-penalty goal contributions. Higher values suggest a higher attacking impact across the entire season.

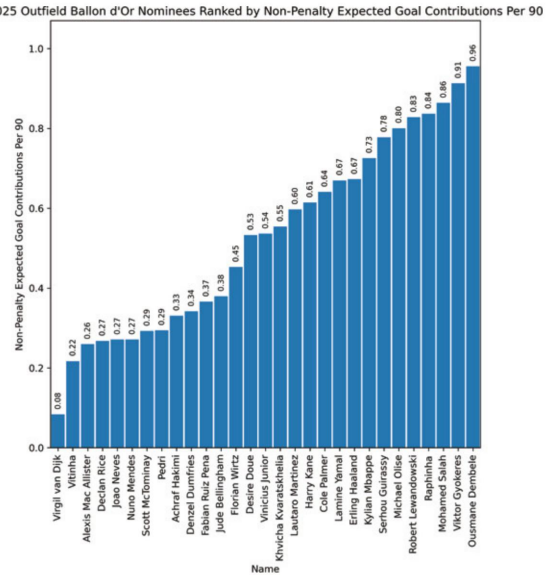


Fig. 2 | 2024-2025 outfield nominees ranked by expected non-penalty goal contributions per 90 minutes. Higher values suggest more efficient attacking productivity.

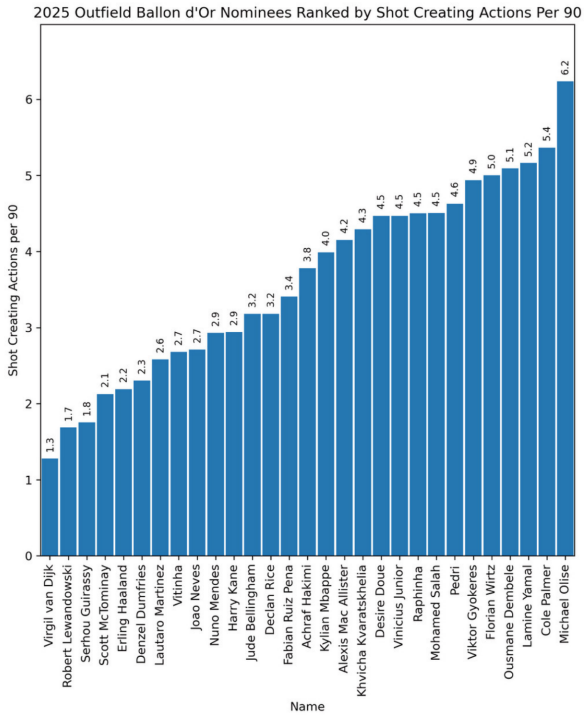


Fig. 3 | 2024-2025 outfield nominees ranked by shot creating actions per 90 minutes.

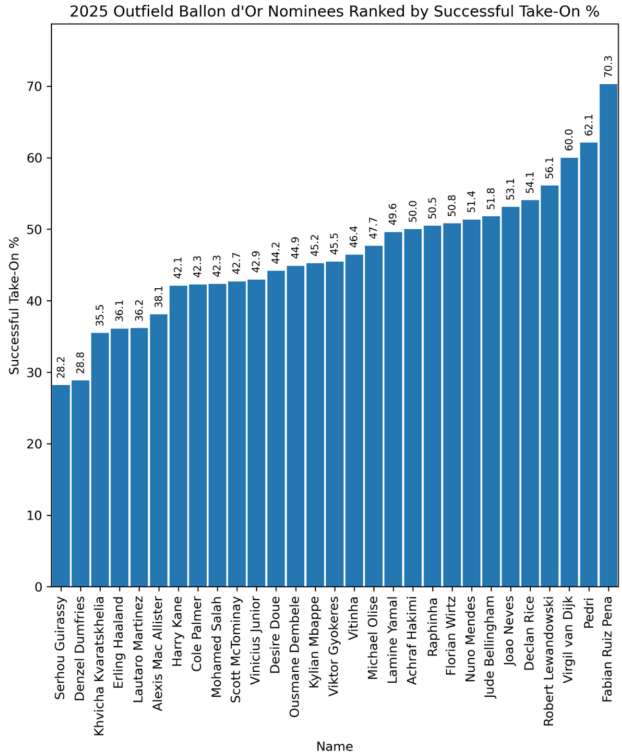


Fig. 4 | 2024-2025 outfield nominees ranked by successful take-on percentage.

3.5 2024-25 Descriptive Visualization Analysis

As expected, attackers dominate both the non-penalty expected goal contribution and non-penalty expected goal contribution per-90 graphs, with Mohamed Salah, Raphinha, and Viktor Gyokeres all occupying spots within the top 4 for both graphs. Interestingly, Ousmane Dembele goes from 7th to 1st when changing from the total to the per-90 regime, while Kylian Mbappe suffers from the opposite problem. With respect to the shot creating actions per-90 graph, high usage wingers dominate the top spots, with the occasional creative central midfielder or striker appearing. Conversely, the successful take-on % graph is much more diverse in positional terms, seemingly accomplishing its goal of skewing less towards attackers. Notably, Virgil Van Dijk, a center back, ranks at 3rd, while Declan Rice, a defensive midfielder, sits at 5th. The top 2, Fabian Ruiz Pena and Pedri, are both also central midfielders.

4. What Past Winners Look Like: Correlation and Similarity

4.1 Feature-Vote Relationships (Correlation Ranking)

To determine the profile of a past Ballon d'Or winner, we first had to gain an idea of the factors that have influenced voting in the past. To do so, we used correlation coefficients to rank features by the strength of their association with within-season vote share. While correlation coefficients hold less statistical power and are generally more susceptible to outliers when used with small datasets, our sample of 84 previous nominees is large enough to draw general conclusions from. Correlation coefficients were calculated using Pearson's correlation:

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Here, x_i and y_i denote the ith observations of the feature and vote share (within a fixed season), *x-bar* and *y-bar* are sample means, σ_x and σ_y are sample standard deviations, and n is the number of nominees in that season.

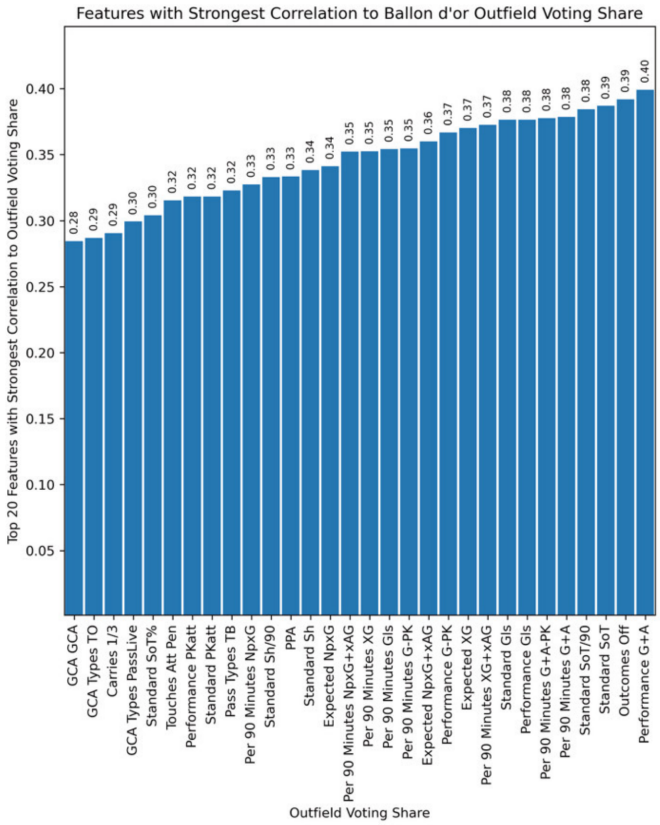


Fig. 5 | Features ranked by strength of their relationship with outfield voting share. Higher values suggest a stronger relationship.

4.2 Building a "Previous Winner Profile"

Using the statistics most strongly correlated with voting outcomes, we can construct a winner profile by summarizing how past winners tend to perform on these features. To ensure a complete winners profile, the selected features were not just the eight most correlated features. Rather, they were chosen to represent as many different areas of play as possible. The final chosen list of features (Goal Contributions, Shots on Target, Offsides, Passes into Opponent Penalty Area, Through Balls, Touches in Opponent Penalty Area, Open Play Passes, Expected Non-Penalty Goal Contributions) all have high correlations with vote share (> 0.3), and collectively represent metrics quantifying passing, shooting, positioning, and offensive output.

After choosing our features, we construct a "past-winner profile" vector by first converting each winner's feature values into within-season percentile ranks (relative to that season's nominee peers), and then averaging those percentile ranks across all the three previous Ballon d'Or winners in our dataset. Percentile ranks were chosen as the method of normalization because they are robust to outliers and quantify the dominance shown compared to peers within their season. We can therefore compute the vector **x bar**, representing the per-feature, peer-relative, average profile of past Ballon d'Or winners, with the following formula:

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\pi}_{s_i}(\mathbf{x}_{w_i}),$$

where $\boldsymbol{\pi}_{s_i}(\mathbf{x}_{w_i})$ denotes the vector of percentile ranks for winner w_i in season s_i (computed relative to that season's nominee peers).

4.3 Nominee Similarity to Past-Winner Profile

To quantify how similar a nominee is to past winners, we compare the nominee's eight-feature vector to the past-winner profile vector. We can measure profile similarity using cosine similarity, which compares vectors by direction, but not magnitude. In this context, cosine similarity lets us compare nominees' relative patterns across the eight features (i.e., which features they are stronger or weaker on), while being less sensitive to overall "scale" differences. A player's cosine similarity to the past-winner profile vector (**x bar**) can be computed using the following formula:

$$\text{cos_sim}(\boldsymbol{\pi}_s(\mathbf{x}_i), \bar{\mathbf{x}}) = \frac{\boldsymbol{\pi}_s(\mathbf{x}_i)^\top \bar{\mathbf{x}}}{\|\boldsymbol{\pi}_s(\mathbf{x}_i)\|\,\|\bar{\mathbf{x}}\|}.$$

Cosine similarity lies within [−1, 1]. In our setting, because the percentile-rank feature vectors are non-negative, cosine similarity lies within [0, 1], where 1 indicates identical feature-direction patterns and 0 indicates orthogonality (no alignment).

4.4 Case Study: Top Contenders Comparison

With methodology established, we can now quantify players' similarity to past Ballon d'Or winners. As a case study, we selected the four main Ballon d'Or "favourites" from the 2024-2025 season (Lamine Yamal, Mohamed Salah, Ousmane Dembele, Raphinha) [3]. These four players were widely viewed as the leading candidates for the award by the public, with all four players producing historic seasons. The spider chart below visualizes each of these favorites' profiles, alongside the past-winners profile calculated in Section 4.2.

We can also plot their cosine similarity scores to the past-winner profile, visualizing which of the favorites' archetype is closest to the winners from previous years.

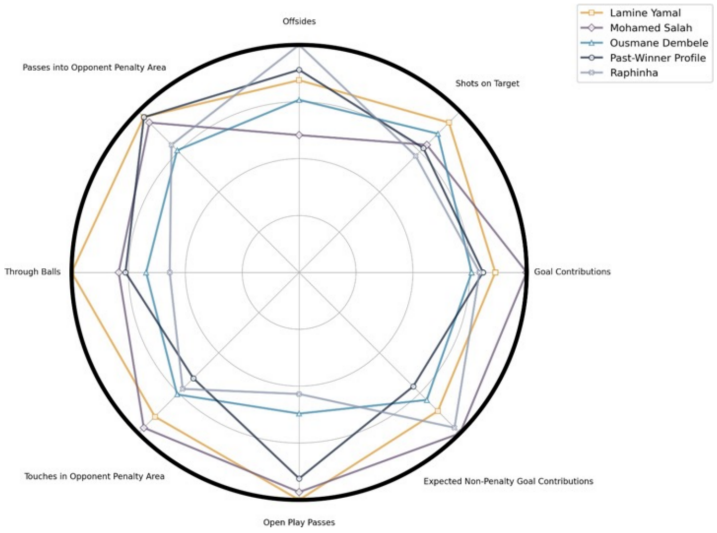


Figure 6 | The four main 2024-2025 Ballon d'Or favourites plotted against the past-winner profile.

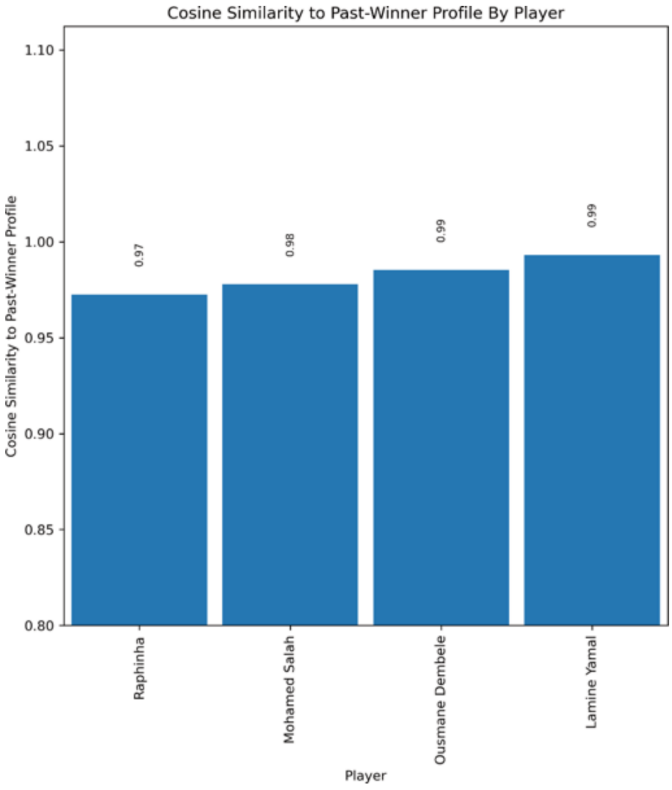


Fig. 7 | The four main 2024-2025 Ballon d'Or favourites ranked by cosine similarity to the past-winner profile.

5. Predicting Ballon d'Or Outcomes with a Nominee-Only Model

This section builds and evaluates a nominee-only predictive model of Ballon d'Or outcomes. The modeling goal is intentionally narrow: using only information observable from season performance and major trophy indicators, we predict within-season vote shares among outfield nominees and assess whether the model can (i) recover the overall ranking structure and (ii) identify the winner and close contenders. We treat this model as a struc-

tured, reproducible benchmark for the "deserved winner" debate: it formalizes a mapping from performance/trophies to voting outcomes, and it provides a way to quantify uncertainty in how close the top of the ballot was. This model is not a claim about objective player value. It is a structured approximation of historical Ballon d'Or voting behavior conditional on the performance and honors signals we observe.

5.1. Target Definition (Points vs Vote Share)

Ballon d'Or voting is recorded as total voting points. Because point totals are not directly comparable across years (e.g., changes in electorate size or voting rules) and because our model is trained across seasons, we define the primary target as outfield vote share within each season:

$$y_{i,s} = \frac{\text{Points}_{i,s}}{\sum_{j \in N_s} \text{Points}_{j,s}},$$

where i indexes an outfield nominee and N_s is the outfield nominee set for season s. This transformation produces a stable unit-free target in [0, 1] that sums to 1 within each season. It also aligns with the substantive question of "how much of the ballot" each nominee captured, and it enables cross-season training without requiring assumptions about point-scale comparability.

We restrict the modeling dataset to outfield nominees only. Goalkeeper voting is governed by different positional priors and feature relevance; mixing keepers and outfielders substantially increases model misspecification in a small-sample setting.

5.2. Feature Sets and Model Choices

Design principles. The dataset is small (four seasons with roughly 27–29 outfield nominees per season; ≈ 114 total rows). We therefore use a disciplined feature design and a model class that is robust to multicollinearity and overfitting. In particular, many football statistics are strongly correlated (e.g., different measures of chance creation, volume vs per-90 rates, minutes vs starts). Feature explosion can produce unstable coefficients and misleading inferences.

Model class: standardized Ridge regression. We use Ridge regression (L2-regularized linear regression) on standardized features. Ridge is well-suited to this setting: it shrinks coefficients smoothly under correlation and provides a stable linear scoring function that can be interpreted and stress-tested. This proportional shrinkage is preferable to alternatives like Lasso or Elastic Net, which tend to arbitrarily zero out highly correlated but contextually important metrics—such as actual goals versus expected goals—when we instead want to preserve the player's complete statistical profile. Let $x_{i,s}$ be a feature vector. The model produces a real-valued score:

$$\hat{z}_{i,s} = \beta_0 + \beta^\top \tilde{x}_{i,s},$$

where x-tilde denotes standardized features.

From scores to vote shares (within-season softmax). Raw linear scores are unbounded and may be negative; vote shares must be nonnegative and sum to 1 within a season. We convert scores to predicted shares using a softmax normalization within each season:

$$\hat{p}_{i,s}(T) = \frac{\exp(\hat{z}_{i,s}/T)}{\sum_{j \in N_s} \exp(\hat{z}_{j,s}/T)},$$

where $T > 0$ is a temperature parameter. Importantly, for fixed T, the softmax transformation is monotonic in z-hat, so the within-season ranking implied by p-hat matches the ranking implied by z-hat. This allows us to evaluate both rank quality (e.g., Spearman) and share-level calibration/residual magnitudes on a common scale.

Feature specifications. We evaluate several closely related feature specifications to understand which modeling choices matter most:

- *Performance-only (core):* a small mixed set combining attacking efficiency (per-90), cre- ation/progression, and durability (minutes/starts).
- *Performance-only (all per-90 + durability):* resolves a common pitfall by expressing creat- ive/progression totals as per-90 rates while keeping explicit durability variables. This is intended to reduce the confounding of "quality" with "minutes played."
- *Performance-only (minimal durability):* a lighter durability variant.
- *Performance-only (log-target):* same disciplined per-90+durance features, but fits Ridge on $\log(y + \epsilon)$ (with very small ϵ) to emphasize relative differences among low-share nominees; predicted shares are still obtained via softmax.
- *Performance-only (simple metric + minutes):* a two-feature "sanity check" model using a single per-90 attacking metric and minutes.
- *With trophies:* the core performance set plus binary indicators for major trophies (league title, UEFA Champions League, domestic cup, major international continental trophy, World Cup). This specification approximates "voter behavior" by allowing a separate trophy signal beyond performance statistics. Because trophies are partly driven by team strength and context (and are therefore endogenous), we treat them as proxies for how voters reward honors rather than as causal drivers of individual quality.

Interpretability outputs. For transparency, we save standardized coefficient plots and permutation- importance plots (computed on the validation season) to illustrate which covariates drive the model's predictions. These diagnostics are treated as qualitative (small n, correlated inputs) rather than as definitive causal statements.

5.3. Train/Validation/Test Design Across Seasons

Season splits. We use a strict season-based split to avoid leakage:

- **Train**: 2021–2022 and 2022–2023
- **Validation:** 2023–2024
- **Test:** 2024–2025

This design matches the real forecasting setting: the model is trained on past seasons and evaluated on a future season.

Regularization selection. For each specification we tune the Ridge penalty α on the validation season. We choose α by maximizing Spearman rank correlation between actual vote shares and predicted shares on 2023–2024. (Operationally, for each α we fit on training seasons, predict validation scores, transform to within-season shares via softmax, and compute Spearman.)

Temperature calibration. While ranking metrics are invariant to monotone transformations, share magnitudes affect residual-based analyses (e.g., "over/under-rated") and "how close was it?" gaps at the top. Therefore, after selecting α, we calibrate the softmax temperature on the validation season by minimizing mean squared error between predicted shares and true shares on 2023–2024. We use MSE for interpretability in absolute vote-share differences (useful for residuals and "how close was it?" gaps), rather than a probabilistic scoring rule like KL or log-loss. This keeps or-

dering fixed but improves share-level interpretability.

Final evaluation. After selecting α and T on 2023–2024, we refit the model on train+validation seasons and evaluate once on 2024–2025.

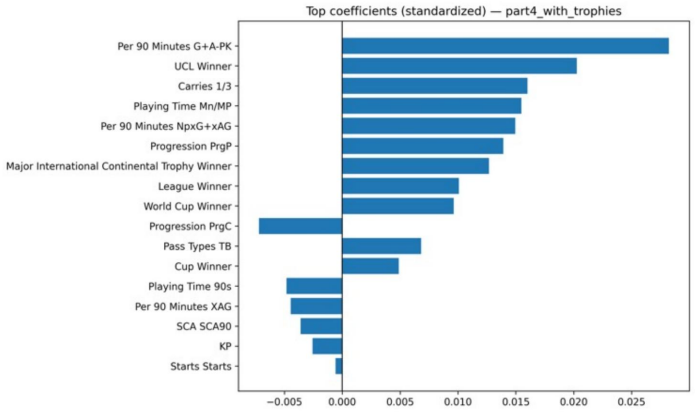


Fig. 8 | Standardized Ridge coefficients for the with_trophies model (validation-trained). Trophy indicators and attacking efficiency/progression features receive the largest absolute weights, consistent with a combined performance-and-honors voting story.

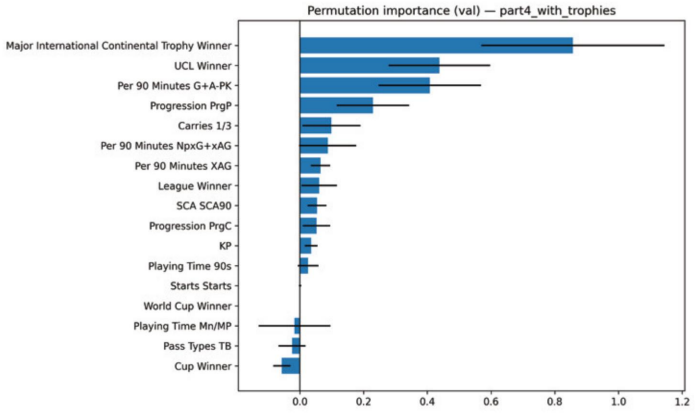


Fig. 9 | Permutation importance on the validation season for the with_trophies model (sanity check). The strongest degradations arise from trophy indicators and core attacking efficiency features, suggesting predictive reliance on these variables.

5.4 Results: Predicted vs. Actual Rankings

Evaluation metrics. We report:

• *Spearman* ρ between actual vote shares and predicted vote shares (ranking fidelity).

• *Pairwise accuracy:* the fraction of nominee pairs ordered correctly by the prediction (ties ignored). This is a robust, human-interpretable measure: "how often do we rank two nominees in the right order?"

• *Winner rank / top-k hit rates:* whether the actual winner is predicted at rank 1 (top-1), within top 3, and within top 5.

Our main finding is that trophies meaningfully improve predictive ranking performance. Across both validation (2023–2024) and test (2024–2025), the with_trophies model produces the strongest overall rank performance and successfully identifies the winner on the held-out test season. On the validation season (2023–2024), with_trophies achieves Spearman ρ = 0.677 and pairwise accuracy = 0.762, and it predicts the winner (Rodri) at rank 4 (top-5 hit). On the test season (2024–2025), with_trophies achieves Spearman ρ = 0.573 and pairwise accuracy = 0.705, and it predicts the winner (Ousmane Dembélé) at rank 1 (top-1 / top-3 / top-5 all hit).

Baselines and ablations. A single-metric baseline using Per 90 Minutes NpxG+xAG is competitive on 2024–2025 in Spearman (ρ = 0.297) but does not win overall and does not consistently identify the winner across seasons. The "similarity-profile" baseline (cosine similarity to an average prior-winner profile) underperforms the main model on both seasons. Among performance-only variants, the disciplined per-90 + durability model (performance_only_all_per90_durability) correctly predicts the 2024–2025 winner at rank 1 but with lower overall rank fidelity than with_trophies (test Spearman ρ = 0.177, pairwise accuracy = 0.559). The log-target variant performs well on validation (Spearman ρ = 0.512) but degrades on test (ρ = 0.062), consistent with instability under small-sample tuning and distribution shift.

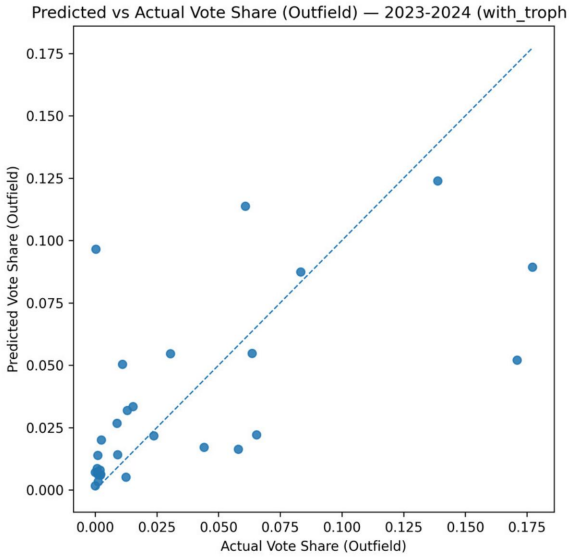


Fig. 10 | Predicted vs actual outfield vote share on the validation season (2023–2024) for the with_trophies model. The dashed line indicates perfect calibration.

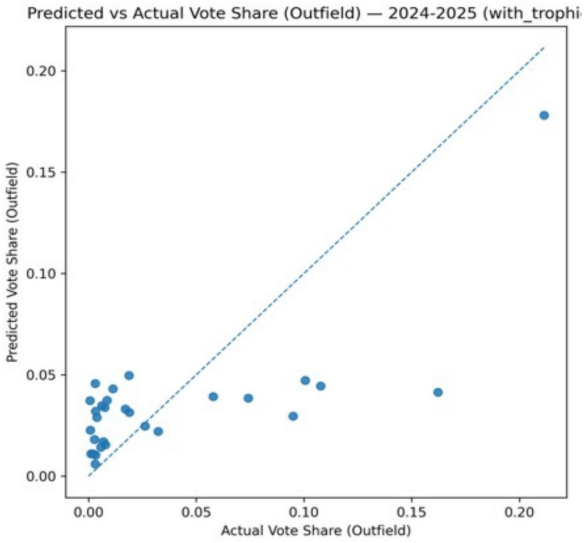


Fig. 11 | Predicted vs actual outfield vote share on the test season (2024–2025) for the with_trophies model. The model is strongest as a rank predictor; share magnitudes remain conservative relative to the most top-heavy realized vote outcomes.

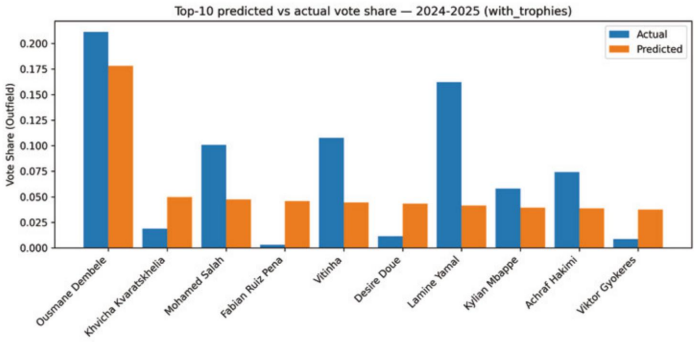


Fig. 12 | Top-10 nominees by predicted share (2024–2025, with_trophies). While the model successfully identifies the top candidate (Dembele) first, it significantly compresses the vote distribution relative to the realized ballot. Specifically, the model heavily under-predicts the actual vote share for major runners-up like Lamine Yamal and Vitinha, while over-predicting the shares of players who received negligible actual votes (e.g., Fabian Ruiz Pena and Desire Doue).

5.5 Over/Under-Rated Analysis (Residuals)

To characterize systematic disagreement between the model and the electorate, we compute residuals in vote-share space:

$$r_{i,s} = \hat{p}_{i,s} - y_{i,s}.$$

Positive residuals indicate nominees the model rates higher than the ballot ("over-rated" by the model); negative residuals indicate nominees the model rates lower than the ballot ("under-rated" by the model). Residual interpretation should be handled cautiously: they may reflect (i) un- measured football contributions, (ii) narrative effects (media momentum, awards discourse), (iii) positional/role factors not captured by our features, and (iv) model misspecification or calibration limits.

In 2024–2025, the with_trophies model predicts the correct winner at rank 1 and produces a comparatively modest residual for Dembélé (−0.03), indicating reasonable calibration at the very top. The largest negative residuals instead belong to other high-vote nominees, most notably Lamine Yamal (−0.12), Raphinha (−0.07), and Vitinha (−0.06), suggesting that the model systematically underestimates the share captured by the broader top tier. This pattern is consistent with a general limitation of small-sample, cross-season models: the realized vote distribution concentrates more sharply among a handful of favourites than the score-to-share mapping learned from prior seasons anticipates.

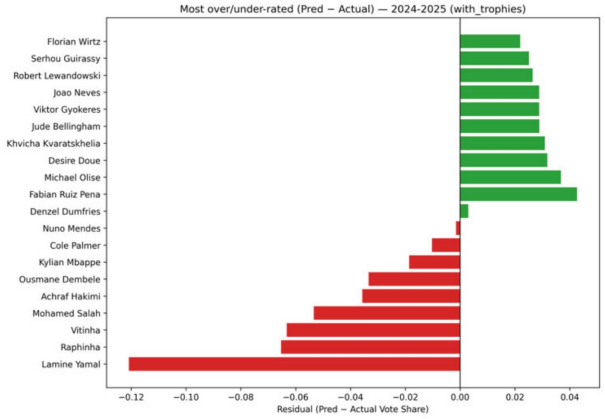


Fig. 13 | Largest residuals (Predicted minus Actual outfield vote share) for 2024–2025 under the with_trophies model. Bars to the right indicate nominees the model rates higher than the ballot; bars to the left indicate nominees the model rates lower than the ballot.

5.6 Robustness/Uncertainty

Bootstrap setup. To quantify uncertainty in predicted shares and, crucially, in "how close was it at the top?", we apply a stratified bootstrap with 800 iterations. We repeatedly resample the training data within each season (preserving season sizes), refit the model, and re-predict 2024–2025. For each nominee we compute:

- a bootstrap mean predicted share,
- a ≈ 90% interval using the 5th and 95th percentiles,
- and probabilities of finishing rank 1 / top 3 / top 5 under the model.

We additionally compute season-level closeness statistics from each bootstrap replicate:

- Top-1 gap: predicted share of rank-1 minus predicted share of rank-2.
- Winner gap: predicted share of the actual winner minus the best predicted share among all other nominees.

Closeness results (2024–2025). Under with_trophies, the predicted race shows clear separation at the top:

- median top-1 gap = 0.1008 share, with 90% interval [0.0178, 0.2148],
- Pr(top-1 gap > 0.02) = 0.93,
- median winner gap = 0.1008 share, with 90% interval [0.0126, 0.2148],
- Pr(actual winner ranked #1 by model) = 0.974.

The 90% interval for the winner gap lies entirely above zero, indicating that under the model's uncertainty, Dembélé's predicted first-place finish is robust across resampled training sets. This supports an interpretation of a dominant predicted winner under the trophy-augmented model, even though the model's predicted share scale remains conservative relative to the realized ballot.

For comparison, the best validation-selected performance-only variant in our pipeline performance_only_logtarget produces a substantially different uncertainty story on 2024–2025, including a much lower probability of ranking the actual winner first (≈ 0.0925) and a negative median winner gap. This divergence highlights that uncertainty quantification is model-dependent and should be presented as conditional on a clearly stated specification.

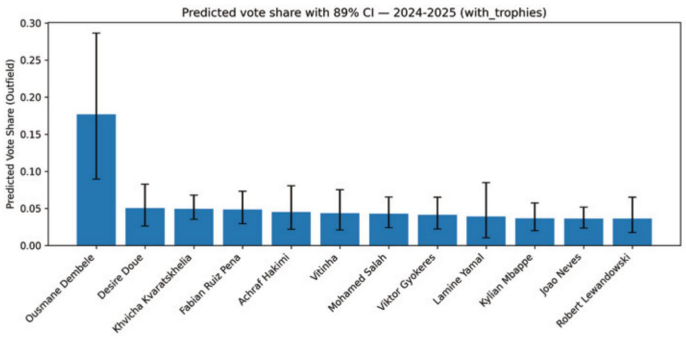


Fig. 14 | Predicted outfield vote shares for 2024–2025 under with_trophies with bootstrap 90% intervals (5th–95th percentiles).

Summary of contributions. This nominee-only model provides a reproducible, season-transferable framework for translating performance (and trophies) into expected vote-share outcomes. In particular, the with_trophies specification achieves the strongest rank performance on both 2023–2024 and 2024–2025 and identifies the winner at rank 1 on the held-out test season. The bootstrap analysis supplies a principled "how close was it?" lens by converting the model into an

uncertainty distribution over the top-of-ballot gap and winner dominance, conditional on the stated modeling assumptions; under with_trophies, that analysis points to clear predicted separation rather than a tight race.

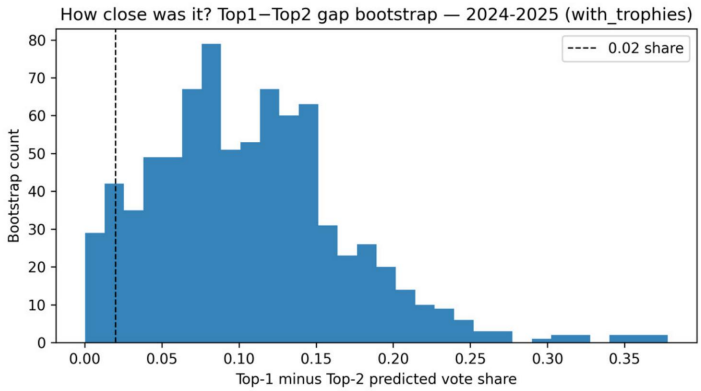


Fig. 15 | Bootstrap distribution of the predicted top-1 minus top-2 vote-share gap for 2024–2025 under with_trophies. The dashed line marks a 0.02 share threshold for interpretability.

6. Synthesis, Conclusions, and Limitations

6.1 Cross-Method Comparison Table

Metric	L. Yamal	M. Salah	O. Dembele	Raphinha
Actual outfield vote share	0.16 (2)	0.10 (3)	0.21 (1)	0.09 (4)
Non-penalty goal contributions per 90	0.67 (4)	0.86 (2)	0.96 (1)	0.84 (3)
Past-winner cosine similarity	0.99 (1)	0.98 (3)	0.99 (2)	0.97 (4)
Predicted outfield vote share	0.04 (3)	0.05 (2)	0.18 (1)	0.03 (4)
Outfield vote shareresidual	-0.12 (4)	-0.05 (2)	-0.03 (1)	-0.07 (3)

Table 1 | Cross-method comparison for top contenders.

6.2 Who Deserved to Win?

The question of "deservedness" depends entirely on the chosen framing of the award. If defined by pure offensive efficiency and output, Ousmane Dembele presents the strongest case, leading the nominees in non-penalty goal contributions per 90 (0.96) and predicted vote share (0.18). However, if the award is viewed as a measure of historical "archetype" alignment, Lamine Yamal holds the narrowest advantage with a cosine similarity score of ≈ 0.99. The model demonstrates that while performance-only metrics are competitive, the inclusion of major trophies is what aligns most closely with realized voting outcomes. Ultimately, Dembele's victory at rank 1 in both the realized ballot and the "with trophies" model suggests he was the most consistent choice under the historical voter reward pattern.

6.3 How Close Was It?

The 2024-2025 race was characterized by a moderate victory by Dembele, with Lamine Yamal also capturing a large portion of the vote share. Our model suggests Dembele was more dominant in "expected" terms than the final realized point totals imply. While Dembele captured a realized vote share of 0.21, the model predicted a similar share of 0.18, resulting in a relatively small residual of −0.03. Bootstrap analysis supports the interpretation of a sizable victory for Dembele: Median Top-1 Gap = 0.1008 with a 90% confidence interval of [0.0178, 0.2148]. Furthermore, the probability that the top-1 gap exceeded a 0.02 share was 0.93, indicating that despite the model's uncertainty, the separation between the top contender was statistically large.

6.4 How Did the Winner Compare to Prior Years?

As shown in both Figure 7 and Table 1, all four selected favorites from the 2024-2025 season (Lamine Yamal, Mohamed Salah, Ousmane Dembele, Raphinha) aligned closely with the past- winner profile under our cosine-similarity measure. Lamine Yamal has the highest similarity score (≈ 0.99), but the differences across the four players are small (≈ 0.02 spread) and should not be over-interpreted. Overall, each of the four favorites appear broadly consistent with the historical winner profile according to this methodology.

6.5 Limitations and Future Work

As we conducted the study, a number of limitations occurred, mainly relating to data and methodology.

Data Errors. While FBref is widely considered a reliable public data source, it is possible that other small, undetected errors were present in the ingested dataset.

Cosine Similarity Limitations. The use of cosine similarity also leads to a limitation in method- ology. Since cosine similarity effectively measures angles between vectors, it does not account for magnitude. This means our methodology will reward those with similar patterns of play relative to their peers with higher similarity scores but cannot account for factors like absolute output. Further analysis would be required to establish a methodology capable of rewarding similar output levels to past winners.

Other Voting Factors. Because of the nature of the data, several key voting factors could not be accounted for. Such factors include narratives, weighted importance of games, and team quality. Further analysis in areas like semantics, detailed tournament-specific statistics, and team quality features could better account for such discrepancies.

6.6 Reproducibility

On January 27, 2026, Opta announced the end of its partnership with FBref. This means that most advanced features gathered for this analysis may stop being available through Engelmann T.'s soccerdata Python package. As such, our data will be published in the project's GitHub repository to aid in reproducibility.

References

[1] Ballon d'Or. (2026). *In Wikipedia*. https://en.wikipedia.org/w/index.php?title=Ballon_d%27Or&oldid=1345115823

[2] Expected goals: What is xG in football and how does it work? (2025, September 26). *BBC Sport.* https://www.bbc.com/sport/football/articles/cgrqd18q0rgo

[3] "*Football transfers, rumours, market values and news*." Retrieved March 28, 2026, from https://www.transfermarkt.com/

Olmeta, L. (2025). *Lucasolmeta/2025-ballon-dor* [Python]. GitHub. https://github.com/lucasolmeta/2025-ballon-dor

Robberechts, P. (2022). *Probberechts/soccerdata* [Python]. GitHub. https://github.com/probberechts/soccerdata

License to the Public

The Authors of this Work grant a CC-BY 4.0 license to the Public.



Illustration by | **Jacqueline Lee '29**