World Cup Player Impact

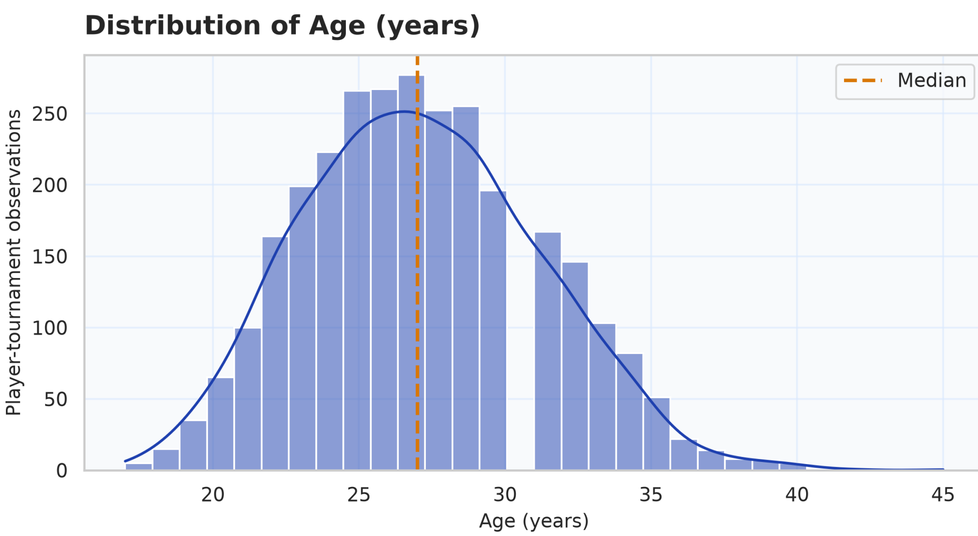
A longitudinal, role-adjusted study of the 2014–2026 tournaments

Abstract

This study analyzes 2,927 player-tournament observations from the 2014, 2018, 2022, 2026 FIFA World Cups to examine combinations of the limited player statistics supplied and identify contributions under-credited by goals and assists. A reliability-adjusted composite combines production, availability, discipline, and recorded positional versatility within tournament-position reference groups.

The best descriptive score-emulation model was Neural Network, with nested player-disjoint RMSE 0.063 and R-squared 0.991. The analysis calls the highest positive benchmark residuals under-credited, not financially undervalued. This terminology is deliberately narrower than market valuation.

The main conclusion is methodological: exposure adjustment, positional context, shrinkage, and uncertainty materially improve interpretation of sparse tournament totals, but no algorithm can reconstruct passing, defending, progression, or goalkeeping actions absent from the data.



Distribution across 2,927 player-tournament observations; dashed line marks the median.

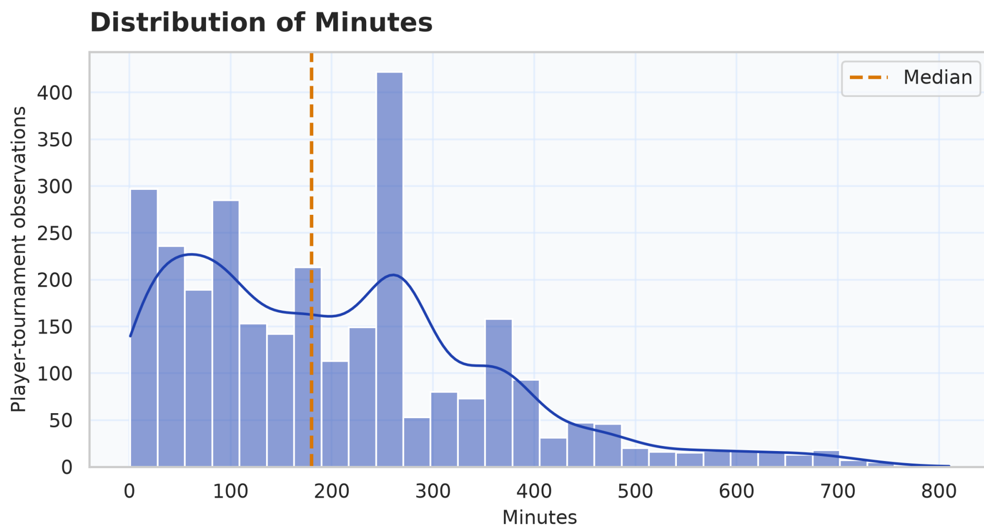
fig_01_distribution_of_age_years. Distribution across 2,927 player-tournament observations; dashed line marks the median.

Introduction

International tournaments create an unusually compressed evaluation setting. Small samples, unequal team progression, substitute appearances, role differences, and low scoring make raw leaderboards unstable. A professional analysis therefore needs explicit exposure, role, and uncertainty controls.

The research question asks which observed combinations characterize impact and whether machine learning can identify players whose broader recorded profile exceeds a traditional production benchmark. The question is answered within the strict information boundary of the supplied CSV files.

This work contributes an auditable score definition, multi-edition validation, model comparison, player-disjoint cross-validation, repeat-player temporal evaluation, explainability, clustering, and a reproducible evidence trail.



Distribution across 2,927 player-tournament observations; dashed line marks the median.

fig_02_distribution_of_minutes. Distribution across 2,927 player-tournament observations; dashed line marks the median.

Literature Review

Football analytics distinguishes event valuation from coarse box-score summaries. Action-value frameworks such as VAEP illustrate why context-rich event streams can measure contributions that goals alone miss; the present data do not contain those events, so this paper adopts a more modest estimand.

Shrinkage is central when rates are estimated from a handful of minutes. Partial pooling toward tournament-position priors reduces the influence of one-match extremes while preserving strong evidence accumulated over longer exposure.

Explainable machine learning is used for error analysis and global dependence rather than causal attribution. Permutation importance, partial dependence, ICE, and SHAP describe fitted model behavior under their assumptions; they do not prove that changing a statistic would cause impact to change.

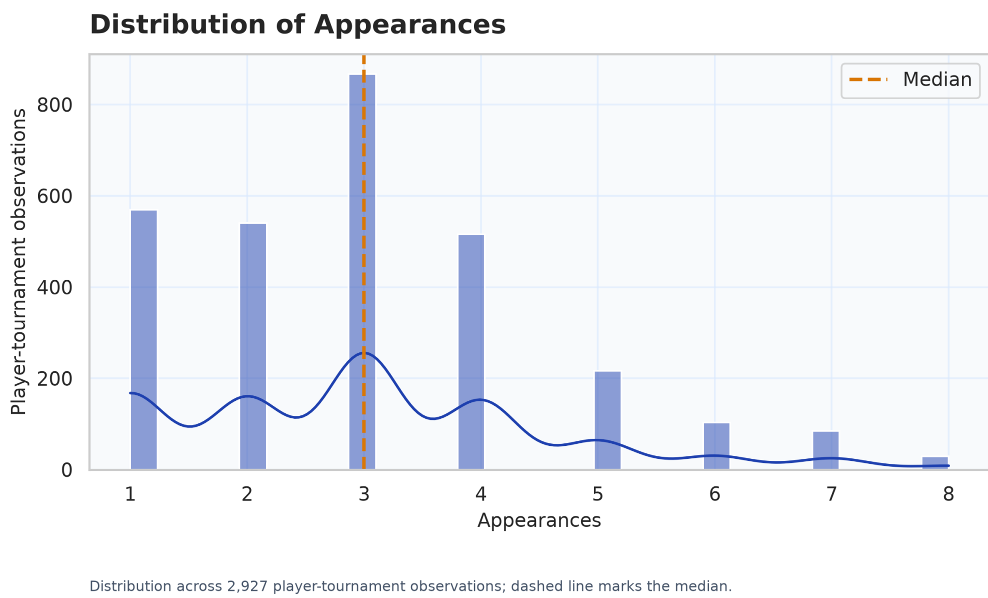


fig_03_distribution_of_appearances. Distribution across 2,927 player-tournament observations; dashed line marks the median.

Research Question and Estimand

The primary estimand is a declared role-relative composite, not latent true ability. Impact is defined at the player-tournament level and therefore combines what happened with how much trusted tournament exposure the player accumulated.

The under-credit estimand is the difference between observed composite impact and an out-of-fold prediction from goals, assists, minutes, position, and tournament. Positive values indicate recorded contribution beyond that benchmark.

For genuine prediction, the longitudinal analysis uses tournament t features to predict the next observed tournament for returning players. This subset contains selection effects because only players returning to a later World Cup are eligible.

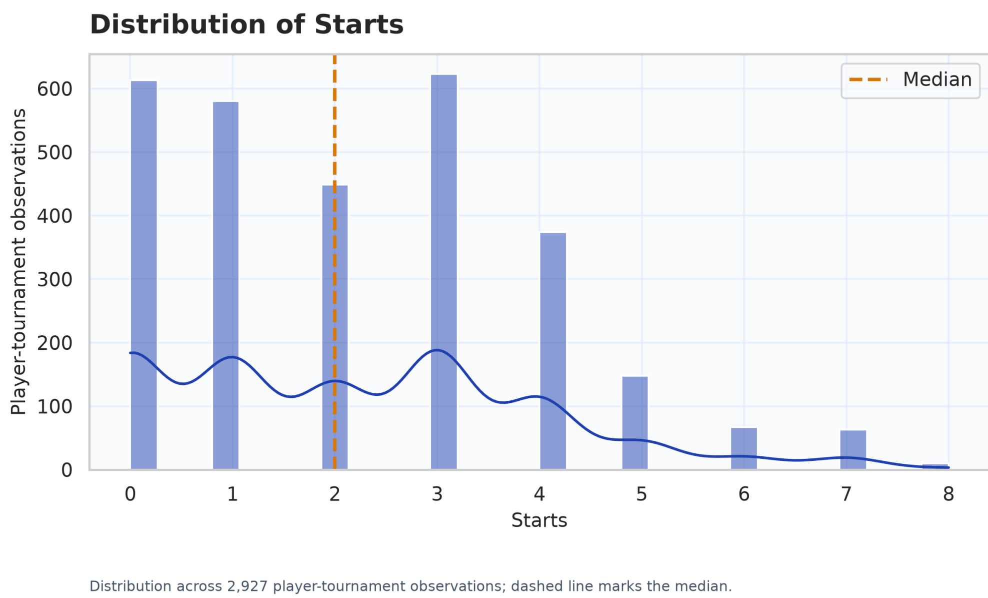


fig_04_distribution_of_starts. Distribution across 2,927 player-tournament observations; dashed line marks the median.

Dataset Description

The harmonized dataset includes 2,927 rows, 2,265 distinct player names, 62 countries, and 521 players observed in multiple editions.

Shared variables cover identity, age, position, squad, appearances, starts, minutes, goals, assists, penalties, cards, player ID, and supplied per-90 rates. Club is available only in later files, while secondary position is often absent.

The player-tournament row is retained rather than aggregating careers because edition context and within-tournament positional normalization are part of the estimand.

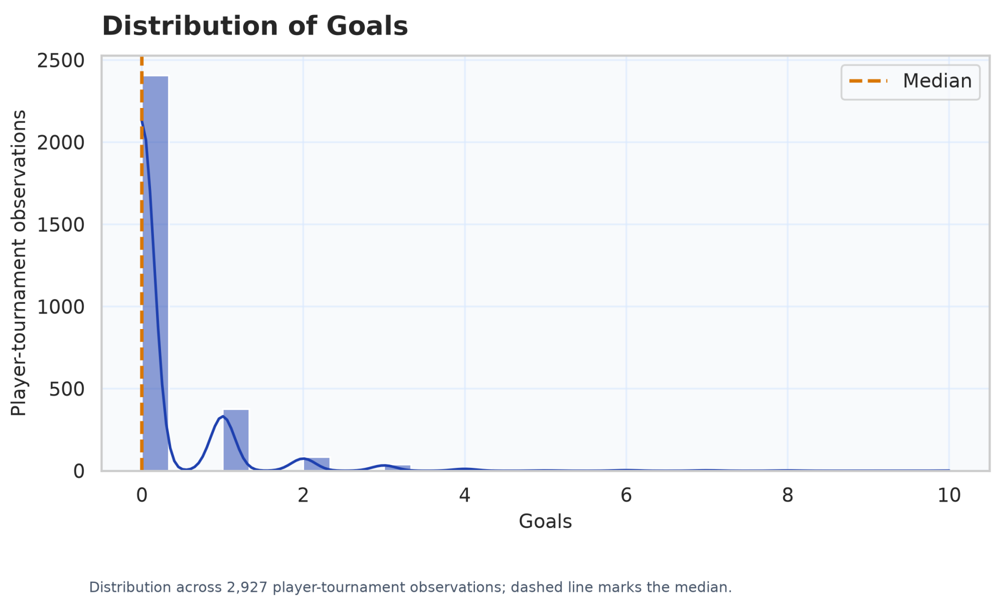


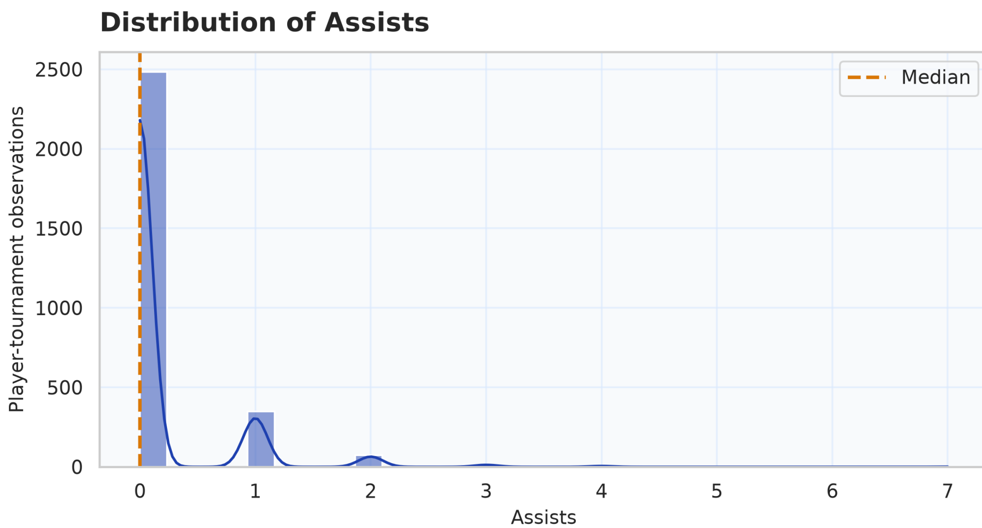
fig_05_distribution_of_goals. Distribution across 2,927 player-tournament observations; dashed line marks the median.

Data Quality and Cleaning

Schema harmonization adds tournament year and a stable observation ID while retaining source columns. Numeric fields are coerced with explicit missingness, categorical absence becomes Unknown, and source files remain immutable.

Validation covers nonnegative counts, starts not exceeding appearances, penalty consistency, total-statistic identities, per-90 rounding, and within-edition player-ID uniqueness. Exact duplicates are removed only when all analytical fields match.

Raw sporting outliers remain in descriptive outputs. Fold-local robust scaling and explicit exposure sensitivity address modelling leverage without rewriting observed tournament records.



Distribution across 2,927 player-tournament observations; dashed line marks the median.

fig_06_distribution_of_assists. Distribution across 2,927 player-tournament observations; dashed line marks the median.

Missing Data Strategy

Age has 1 missing observations. Modelling uses the tournament-position median and a missingness indicator, preventing silent information loss.

Secondary position and club absence are represented as explicit unknown categories because their missingness is structural and edition-dependent. They are not imputed as if a specific club or secondary role had been observed.

Robustness tables report missingness and the chosen strategy. Analyses requiring wholly absent variables are skipped and documented in the capability report.

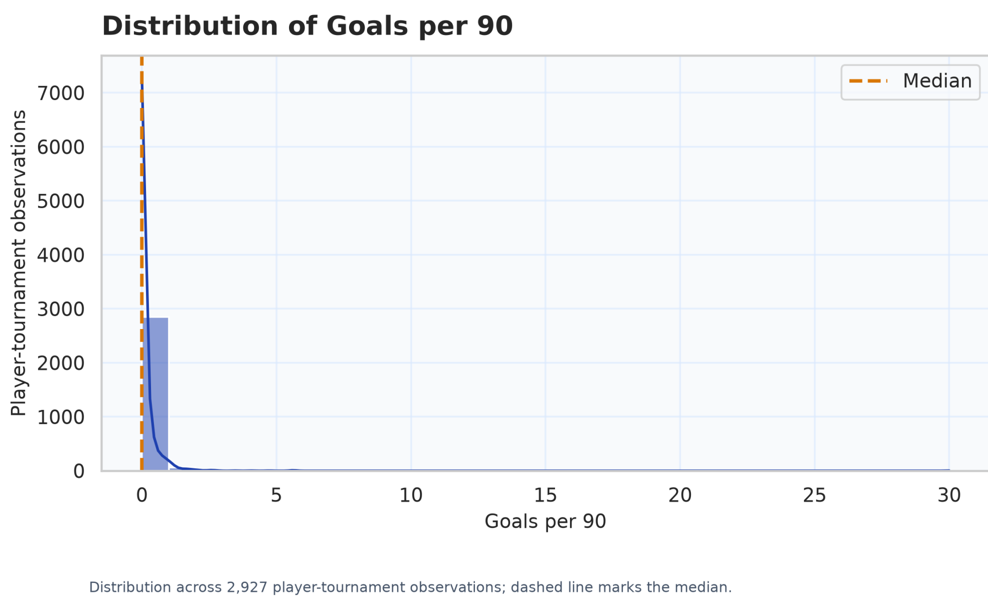


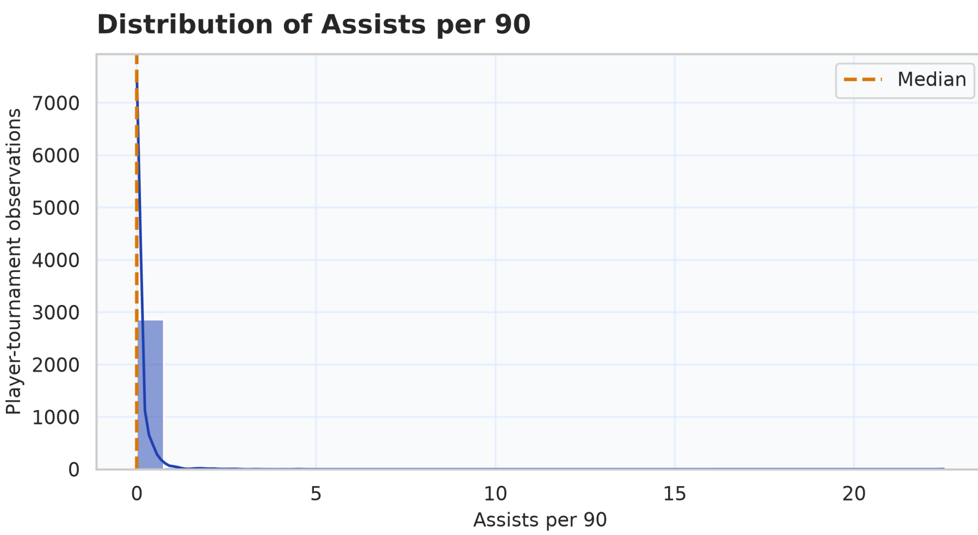
fig_07_distribution_of_goals_per_90. Distribution across 2,927 player-tournament observations; dashed line marks the median.

Feature Engineering

Per-90 rates are recomputed from counts and exact minutes with undefined zero-exposure denominators. Goals excluding penalties, goal contributions, start rate, minutes per appearance, relative squad minutes, discipline burden, and recorded versatility provide interpretable components.

Scoring and assist rates are shrunk toward the exposure-weighted tournament-position mean using a configurable 180-minute prior. A 90-minute cameo therefore receives more pooling than a seven-match tournament.

All equations and interpretations appear in the engineered feature dictionary. Manual polynomial expansion is omitted because kernel, tree, boosting, and neural estimators already test nonlinearity without multiplying unstable sparse-tournament terms.



Distribution across 2,927 player-tournament observations; dashed line marks the median.

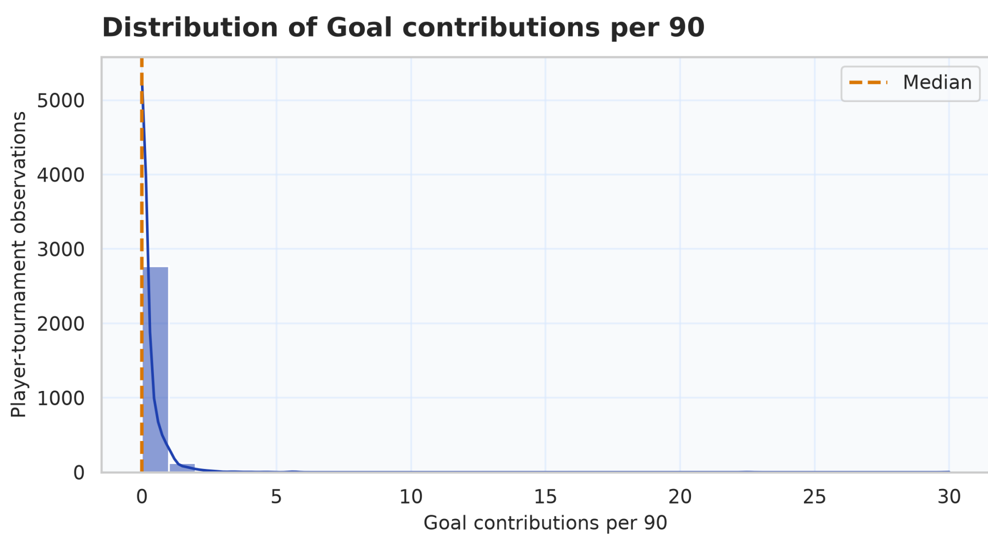
fig_08_distribution_of_assists_per_90. Distribution across 2,927 player-tournament observations; dashed line marks the median.

Methodology

Production receives 45 percent of the declared weight, availability 35 percent, discipline 10 percent, and recorded versatility 10 percent. Components are robust standardized within tournament and primary position before weighting.

Availability is not a pure skill measure: it captures selection, fitness, team progression, and coach trust. Its inclusion reflects tournament contribution rather than counterfactual talent.

Goalkeepers are explicitly flagged. With no saves, shots faced, claims, distribution, or clean sheets, their composite is primarily an availability and discipline description and should not be used as a comprehensive goalkeeper ranking.



Distribution across 2,927 player-tournament observations; dashed line marks the median.

fig_09_distribution_of_goal_contributions_per_90. Distribution across 2,927 player-tournament observations; dashed line marks the median.

Exploratory Analysis

The figure atlas separates raw totals from rate measures and reliability-adjusted estimates. Distributions are zero-inflated for goals and assists, while minutes and starts reflect tournament progression and squad hierarchy.

Position and edition comparisons demonstrate why pooled raw ranks are misleading. Within-position standardization improves comparability but does not erase tactical heterogeneity inside broad defender, midfielder, and forward labels.

All static graphics are rendered at 300 DPI under a consistent colorblind-safe theme. The figure manifest supplies captions, statistical interpretations, football interpretations, data sources, and limitations.

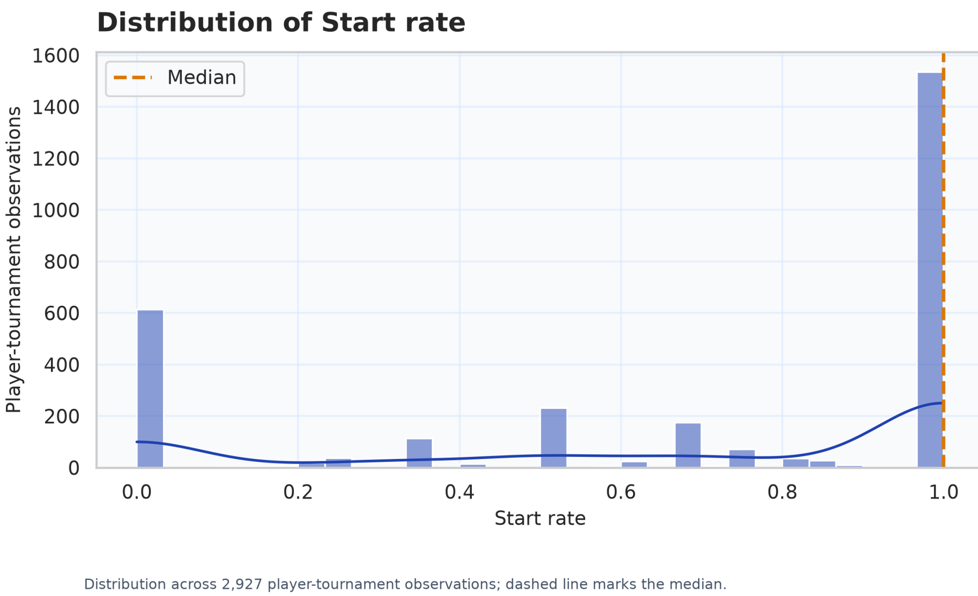


fig_10_distribution_of_start_rate. Distribution across 2,927 player-tournament observations; dashed line marks the median.

Position Analysis

Forwards show the strongest scoring concentration, while midfielders distribute more weight toward assists and defenders toward exposure. These are broad descriptive tendencies rather than fixed tactical archetypes.

Classical ANOVA, Welch ANOVA, Kruskal-Wallis, Levene diagnostics, and pairwise Welch t and Mann-Whitney tests quantify position differences. Effect sizes and adjusted p-values accompany significance tests.

Wingbacks and center backs cannot be isolated reliably because the source uses broad primary and optional broad secondary positions, not detailed roles.

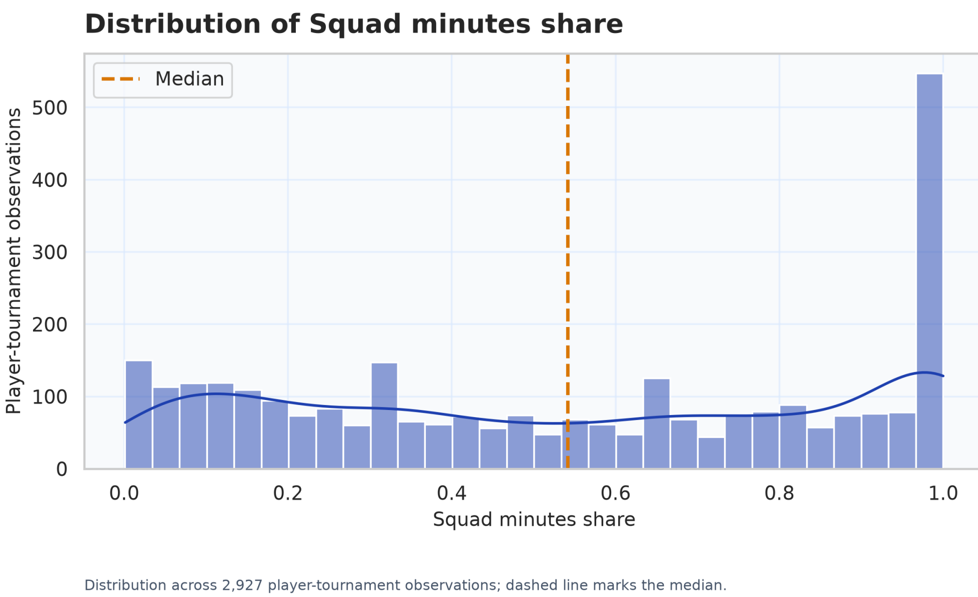


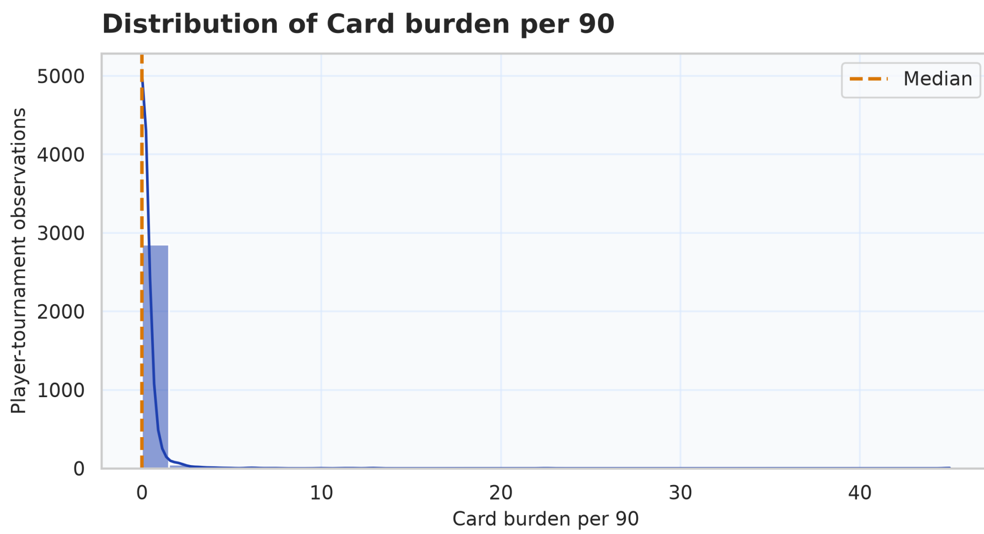
fig_11_distribution_of_squad_minutes_share. Distribution across 2,927 player-tournament observations; dashed line marks the median.

Country Analysis

Country summaries aggregate offensive production, availability, discipline, impact, and consistency for each edition. Relative strengths and weaknesses are defined against teams in the same tournament.

Total minutes strongly reflect progression, so the report keeps exposure and rate metrics separate. Team depth cannot be inferred from a single mean without examining dispersion and player counts.

Defensive performance and passing quality are marked unavailable for every country rather than inferred from cards or goals. Discipline is a conduct dimension, not a defensive-quality proxy.



Distribution across 2,927 player-tournament observations; dashed line marks the median.

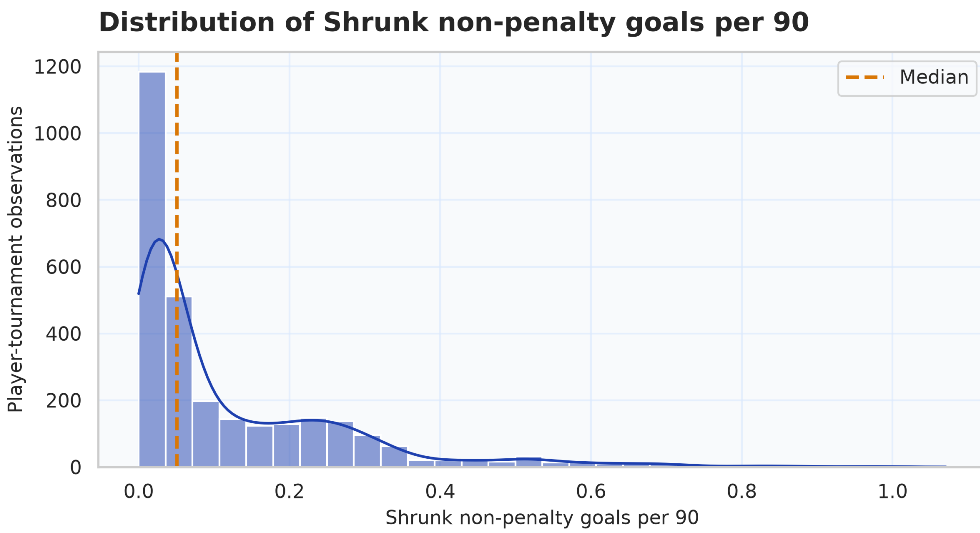
fig_12_distribution_of_card_burden_per_90. Distribution across 2,927 player-tournament observations; dashed line marks the median.

Statistical Inference

Pearson, Spearman, and Kendall coefficients distinguish linear from monotonic association. Shapiro-Wilk and Levene tests diagnose distribution and variance assumptions without automatically invalidating large-sample parametric comparisons.

Omnibus and pairwise tests report effect sizes, confidence intervals, bootstrap intervals, permutation evidence, and Benjamini-Hochberg corrections. Practical interpretation is based on effect magnitude and football context, not p-values alone.

Variance inflation factors identify redundant inputs. High multicollinearity among minutes, starts, and appearances is expected and informs feature selection and interpretation.



Distribution across 2,927 player-tournament observations; dashed line marks the median.

fig_13_distribution_of_shrunk_non_penalty_goals_per_90. Distribution across 2,927 player-tournament observations; dashed line marks the median.

Dimensionality Reduction

PCA provides the primary auditable low-dimensional representation. Scree, cumulative variance, loadings, and biplot data support component interpretation in terms of availability, production, age, discipline, and versatility.

UMAP and t-SNE provide exploratory neighborhood maps with fixed seeds. Their axes have no direct football meaning, and distances—especially global distances—must not be overinterpreted.

Agreement across embeddings is treated as qualitative support for profile structure, not validation of an objective taxonomy.

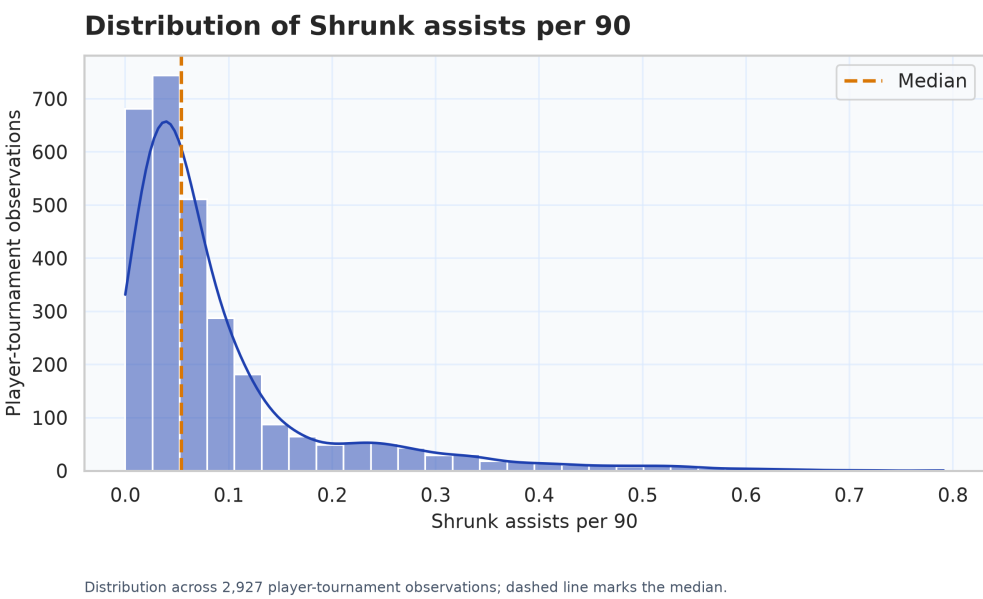


fig_14_distribution_of_shrunk_assists_per_90. Distribution across 2,927 player-tournament observations; dashed line marks the median.

Unsupervised Player Profiles

K-Means, hierarchical clustering, Gaussian mixtures, DBSCAN, HDBSCAN, and spectral clustering are compared when computationally supported. Silhouette, Davies-Bouldin, Calinski-Harabasz, gap statistic, and noise fraction guide selection.

Archetype names are assigned only after examining observed cluster profiles. Permitted descriptions include durable starters, high-production finishers, creative contributors, experienced veterans, and versatile contributors.

The analysis does not label defensive specialists or possession controllers because no tackles, interceptions, passes, or possession events exist.

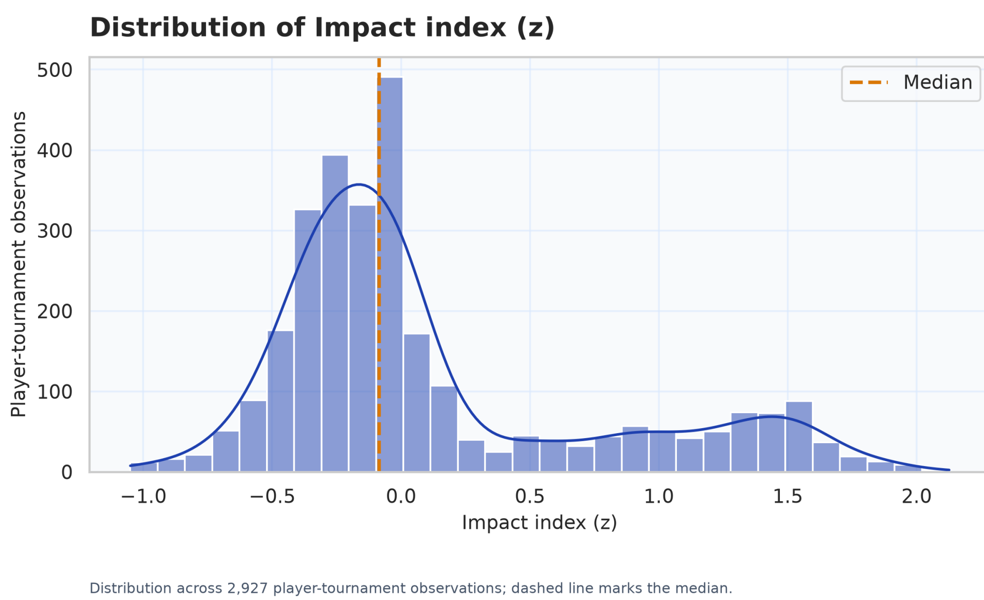


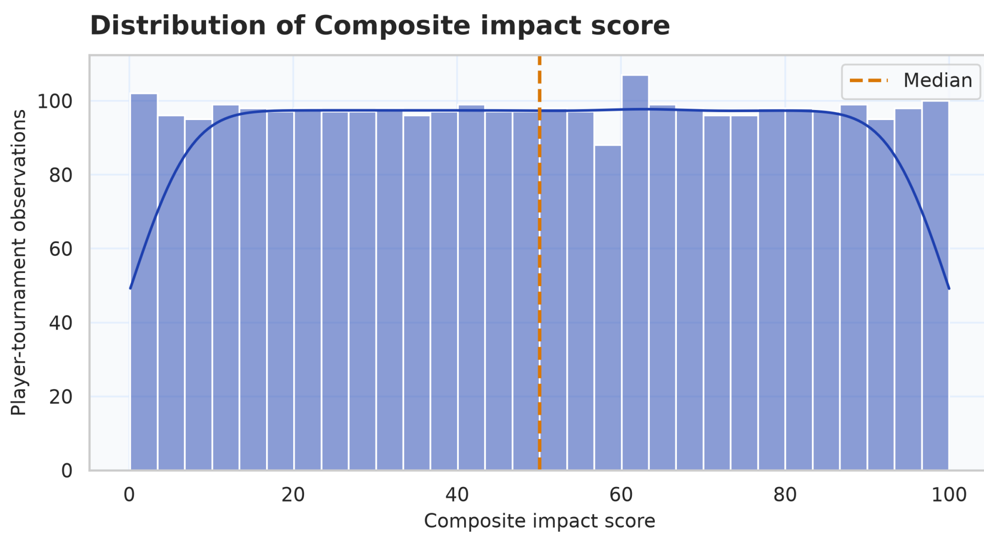
fig_15_distribution_of_impact_index_z. Distribution across 2,927 player-tournament observations; dashed line marks the median.

Predictive Machine Learning

Nine estimator families are capability-gated: linear regression, Elastic Net, random forest, gradient boosting, XGBoost, LightGBM, CatBoost, support-vector regression, and a multilayer perceptron. The best descriptive model is Neural Network.

Optuna tuning is nested inside player-disjoint outer validation. Preprocessing, imputation, encoding, scaling, and tuning use training folds only. Metrics include MAE, RMSE, safe MAPE, R-squared, and adjusted R-squared when defined.

The cross-sectional target is constructed from the inputs, so high accuracy primarily shows how an estimator recovers the declared formula and interactions. It is not evidence of external predictive validity.



Distribution across 2,927 player-tournament observations; dashed line marks the median.

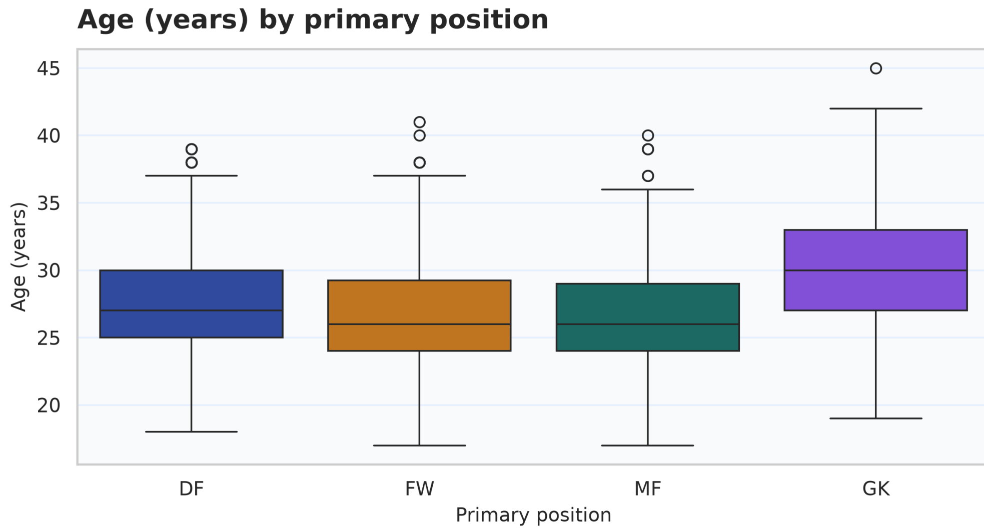
fig_16_distribution_of_composite_impact_score. Distribution across 2,927 player-tournament observations; dashed line marks the median.

Longitudinal Prediction

Repeat-player pairs provide the non-circular predictive evaluation: the current tournament predicts impact at the next observed tournament. Training uses earlier target editions and testing moves forward in time.

Return to a later World Cup is itself selective, and position, tactics, fitness, and team qualification can change between editions. Temporal metrics should therefore be interpreted as persistence among returners.

Wide error relative to descriptive emulation demonstrates the difference between reconstructing a declared score and forecasting future tournament performance.



Boxes show the interquartile range and median; points beyond whiskers remain visible rather than being deleted.

fig_17_age_years_by_primary_position. Boxes show the interquartile range and median; points beyond whiskers remain visible rather than being deleted.

Explainable AI

Permutation importance measures the increase in held-out-style error after shuffling one feature. Partial dependence summarizes mean fitted response, and ICE reveals heterogeneous player-level response patterns.

SHAP is generated when the fitted model and installed explainer are compatible. Local explanations are tied to specific model predictions and never presented as causal decompositions of actual football outcomes.

Explanation concordance is more informative than a single importance chart. Disagreement can reveal correlation, extrapolation, or model-specific behavior.

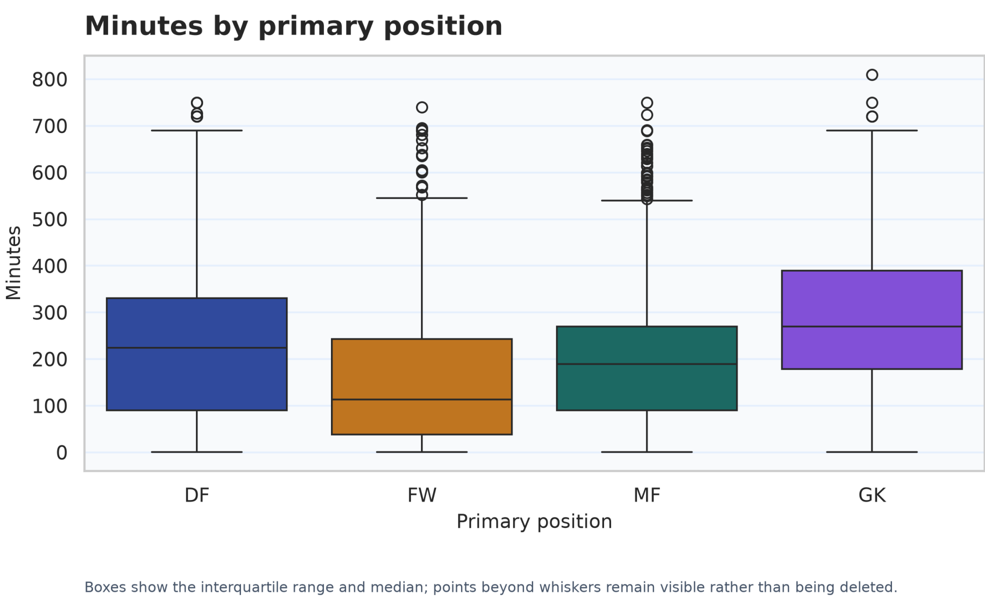


fig_18_minutes_by_primary_position. Boxes show the interquartile range and median; points beyond whiskers remain visible rather than being deleted.

Under-Credit Analysis

The leading positive residual in the generated ranking is Ramin Rezaeian. Each listed player includes goals, assists, minutes, observed impact, traditional prediction, residual, exposure-sensitive interval, and component explanation.

The top 25 positive and negative residual tables are comparative diagnostics. A negative residual means traditional output is high relative to the broader declared index, not that a player performed poorly.

Because no salary, transfer fee, public perception, or market estimate exists, the paper avoids the economic meaning of undervalued and uses under-credited throughout the interpretation.

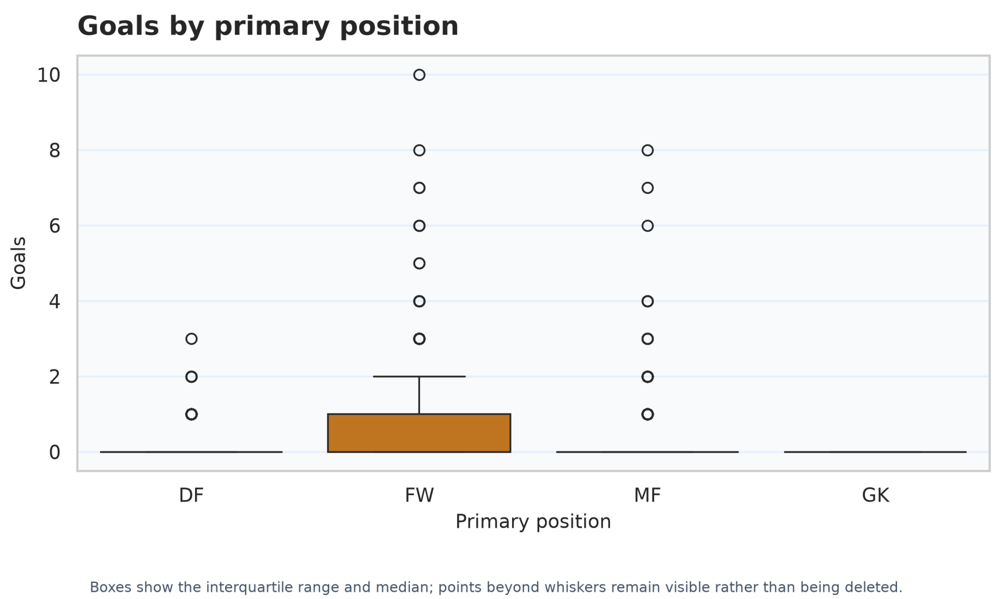


fig_19_goals_by_primary_position. Boxes show the interquartile range and median; points beyond whiskers remain visible rather than being deleted.

Robustness and Sensitivity

Alternative component weights emphasize production, availability, or equal dimensions. Spearman rank agreement and top-25 overlap show how much substantive rankings depend on the declared weighting choice.

Minute thresholds test sensitivity to cameo outliers, while bootstrap intervals and cross-validation dispersion quantify sampling and model uncertainty. Fixed seeds make stochastic methods reproducible; stability checks vary the seed where relevant.

Robustness cannot repair measurement omission. Stable conclusions about a narrow index remain narrow conclusions.

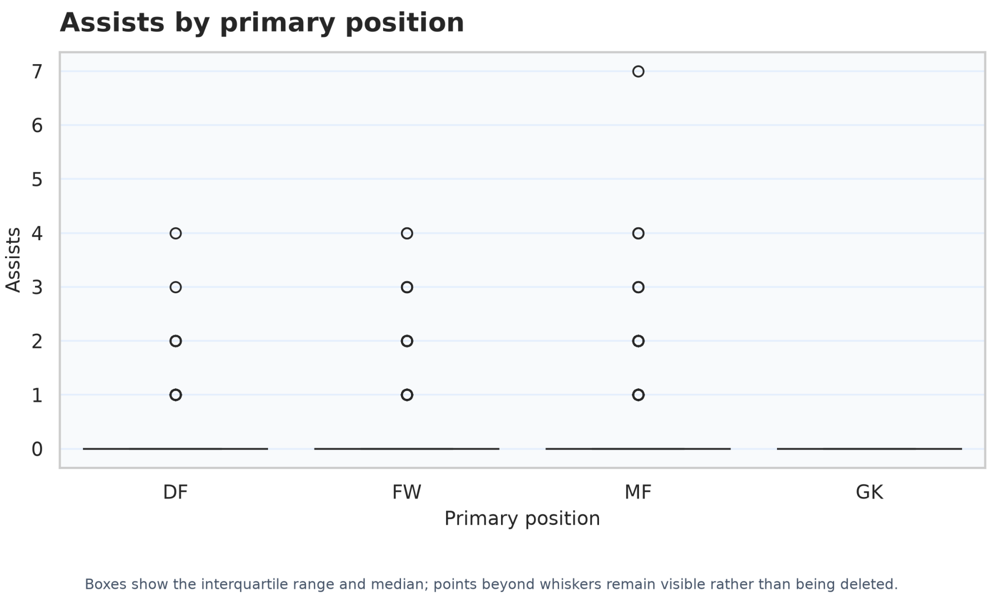


fig_20_assists_by_primary_position. Boxes show the interquartile range and median; points beyond whiskers remain visible rather than being deleted.

Limitations

The dataset lacks passes, shots, expected goals, defensive actions, pressures, possessions, match state, opponent strength, detailed roles, goalkeeper actions, injuries, and economic value. As a result, all-around football impact is only partially observed.

The composite weights are normative, broad positions mask tactical diversity, and availability partly reflects team progression. Multiple rows for returning players require grouped validation and do not form an independent random sample.

Future research should link event data, lineup context, opponent strength, expected-threat or VAEP values, goalkeeper shot-stopping, and market information. A preregistered outcome definition would permit stronger external validation.

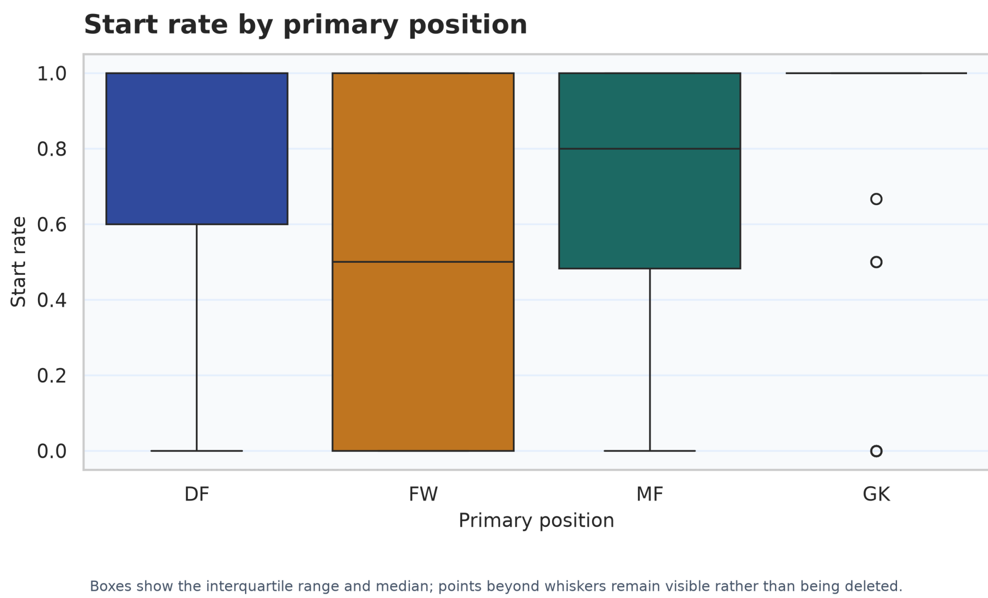


fig_21_start_rate_by_primary_position. Boxes show the interquartile range and median; points beyond whiskers remain visible rather than being deleted.

Conclusion

The observed combination that most consistently characterizes the declared impact index joins reliability-adjusted production with tournament availability, then adds smaller discipline and versatility terms. Role normalization and shrinkage prevent raw scoring totals from dominating every conclusion.

Machine learning identifies statistically under-credited player-tournament performances relative to a traditional benchmark, but the claim is explicitly conditional on the constructed index and supplied variables.

The central portfolio lesson is that professional sports analytics is as much about measurement boundaries, leakage control, uncertainty, and honest language as it is about algorithm breadth.

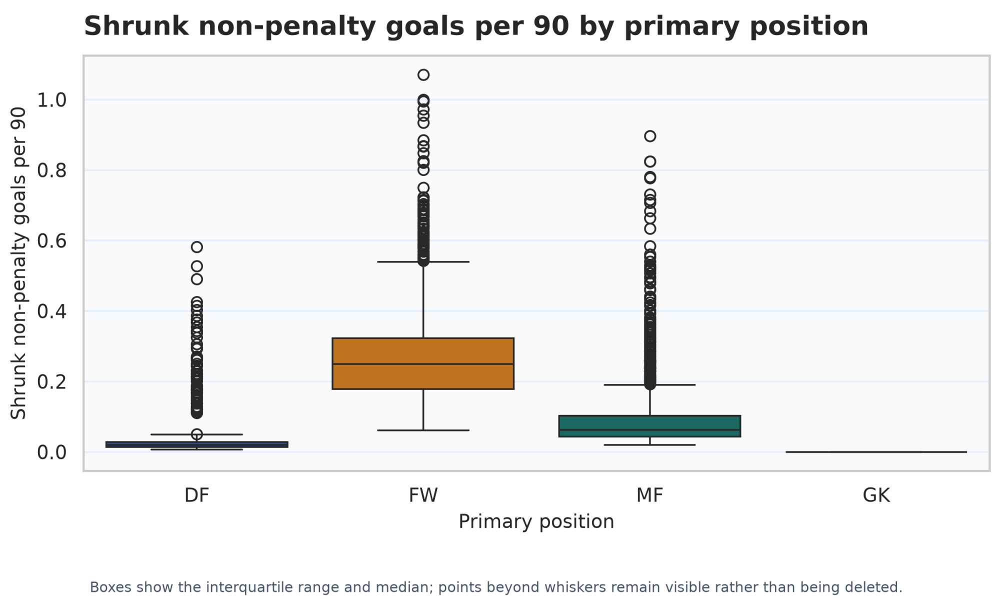


fig_22_shrunk_non_penalty_goals_per_90_by_primary_position. Boxes show the interquartile range and median; points beyond whiskers remain visible rather than being deleted.

References

Decroos, T., Bransen, L., Van Haaren, J., and Davis, J. (2019). Actions Speak Louder than Goals: Valuing Player Actions in Soccer. KDD, 1851–1861. doi:10.1145/3292500.3330758.

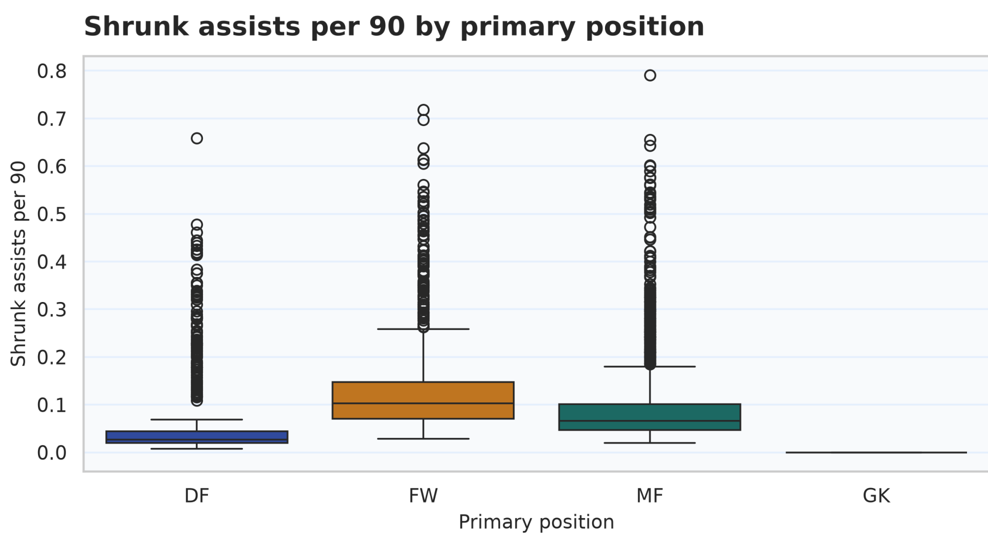
Pappalardo, L. et al. (2019). A public data set of spatio-temporal match events in soccer competitions. Scientific Data 6, 236. doi:10.1038/s41597-019-0247-7.

Lundberg, S. M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. NeurIPS 30.

Breiman, L. (2001). Random Forests. Machine Learning 45, 5–32.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426.

All code, configuration, raw inputs, processed tables, figures, serialized models, and manifests needed for reproduction are included in the repository. The appendix figure atlas references every generated graphic.



Boxes show the interquartile range and median; points beyond whiskers remain visible rather than being deleted.

fig_23_shrunk_assists_per_90_by_primary_position. Boxes show the interquartile range and median; points beyond whiskers remain visible rather than being deleted.