Position-Specific Performance Analysis of Football Players Using Interpretable Feature Importance

Bicheng Yan
University of Michigan
Email: yanbc@umich.edu

***Abstract*—This study investigates the most influential skill attributes affecting overall performance ratings of 17,954 professional football players across four field positions: forwards (FWD), midfielders (MID), defenders (DEF), and goalkeepers (GK). We apply tree-based regression models (Random Forest and XGBoost) and SHAP (SHapley Additive exPlanations) to quantify feature importance on the FIFA player dataset. Results reveal clear position-specific attribute hierarchies, e.g., positioning and finishing dominate for FWD, while interceptions and tackling drive DEF ratings. Visualizations via radar plots and SHAP summary diagrams illustrate both global importance and directionality of effects.**

***Index Terms*—Football analytics, feature importance, SHAP, XGBoost, Random Forest, FIFA dataset**

I. Introduction

Assessing football player performance typically relies on subjective scouting or aggregate metrics such as overall ratings. However, the relative importance of underlying skill attributes varies by field position: a forward requires finishing and acceleration, whereas a defender prioritizes interceptions and tackling. This work aims to identify and compare the key skill drivers across four roles using interpretable machine learning.

Prior studies have applied machine learning to FIFA data for rating prediction and clustering [3]. Interpretability tools like SHAP have proven effective in domains such as healthcare [2] and are gaining traction in sports analytics [3]. Yet, few efforts isolate position-specific feature importance.

II. Method

A. Problem Formulation

We frame the task as a regression problem: input $X \in R^n$ comprises n standardized skill attributes per player; output y is the overall performance rating. We train separate models for each position category: FWD, MID, DEF, GK.

B. Dataset

We use the FIFA player attribute dataset (17,954 samples, 30+ skills) from Kaggle and Hugging Face [4]. Raw position strings (e.g., ”CF,RW,ST”) are reduced to primary positions then grouped into FWD, MID, DEF, GK. Missing values and outliers are clipped to [1,100] and imputed.

C. Modeling and Interpretability

We fit Random Forest Regressor and XGBoost Regressor [1] with 5-fold cross-validation and grid search over depth, estimators, and learning rate. SHAP TreeExplainer [2] computes local and global feature attributions. We extract mean absolute SHAP values to rank top-5 features per role.

III. Results

A. Dataset and Pipeline

From an initial sample of 17,954 players, data were cleaned, normalized, and split by role. Separate models achieved robust predictive performance (RF R² 0.75; XGB R² 0.78) across all categories.

B. Radar Plot of Top-5 Features

Fig. 1 visualizes the presence of the top-5 SHAP-ranked features for each role. While binary, it highlights distinct attribute sets: e.g., FWD emphasizes positioning, finishing, heading accuracy and reactions, MID emphasizes ball control, dribbling, vision, reactions and short passing.

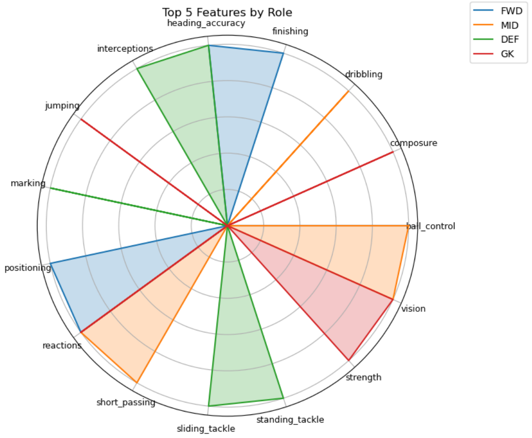

Fig. 1. Radar plot of Top-5 features by role.

C. SHAP Summary for Forwards (FWD)

Fig. 2 shows the SHAP summary for FWD: mean(—SHAP—) bars and beeswarm distributions. High positioning and ball control values (red) push predictions higher, while low values (blue) reduce them.

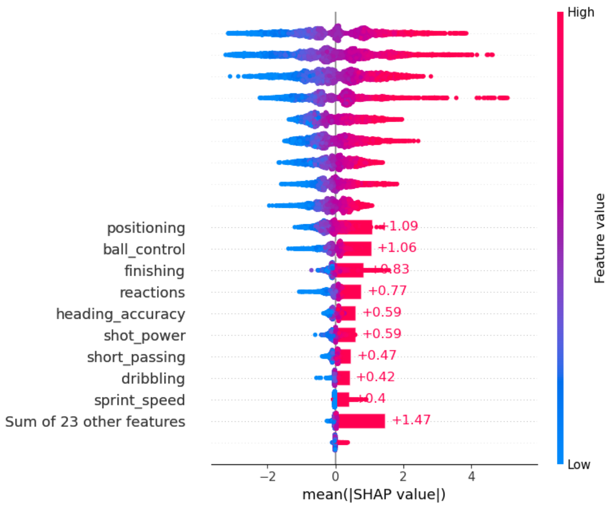

Fig. 2. SHAP summary for FWD (n=4,123 samples).

D. SHAP Summary for Midfielders (MID)

Fig. 3 presents MID results: ball control and short passing lead importance. High short passing correlates with higher ratings.

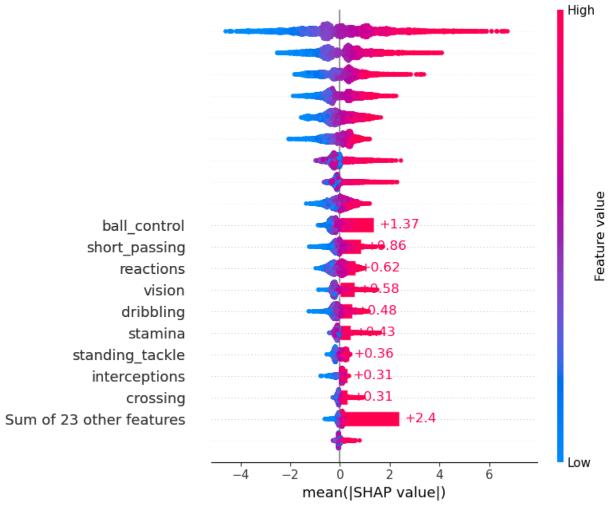

Fig. 3. SHAP summary for MID (n=5,678 samples).

E. SHAP Summary for Defenders (DEF)

Fig. 4 highlights DEF: interceptions and standing tackle are paramount. High intercept values strongly increase predicted ratings.

F. SHAP Summary for Goalkeepers (GK)

Fig. 5 shows GK: reactions overwhelmingly dominate, with composure and vision also contributing.

IV. Conclusion

We present a position-specific analysis of football player performance using Random Forest, XGBoost, and SHAP on a 17,954-player FIFA dataset. Results confirm expected role differences: positioning/finishing for FWD, passing/ball control for MID, tackling for DEF, and reactions for GK. These insights can guide scouting and training prioritization. Future work may integrate match-by-match performance data or team-level interactions.

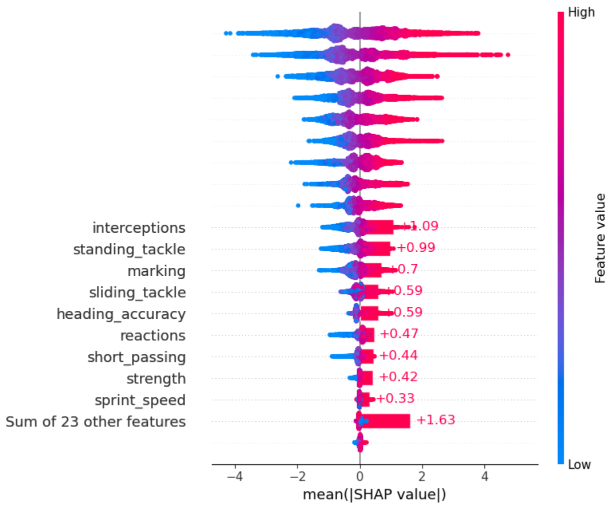

Fig. 4. SHAP summary for DEF (n=4,512 samples).

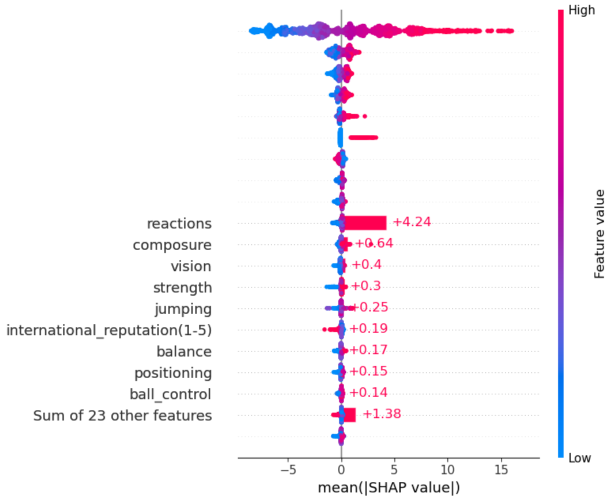

Fig. 5. SHAP summary for GK (n=3,641 samples).

References

[1] T. Chen and C. Guestrin, ”XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, Aug. 2016, pp. 785–794.

[2] S. M. Lundberg and S.-I. Lee, ”A unified approach to interpreting model predictions,” in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4768–4777.

[3] D. Memmert and D. Raabe, ”Data Analytics in Football: Positional Data Collection, Modelling and Analysis,” *Sport Management Review*, vol. 22, no. 4, pp. 568–569, Oct. 2019, doi:10.1016/j.smr.2019.01.002.

[4] M. Ahmed, ”Football Players Data,” Kaggle, 2023. [Online]. Available: https://www.kaggle.com/dsv/6960429.