Abstract of Data-Driven Selection and Role Assignment for Emlyon Business School Football Team

CHAMAKHI Ayoub, CAUHAPE Agathe, PONCET Maude, PURKAYASTHA Aakash

Professor Dr. Franck JAOTOMBO

December 4, 2024

1 Introduction

Emlyon Business School, known for its integration of sports excellence with academic rigor, is gathering top talent to strengthen its soccer club. The challenge lies in identifying the best football talent within the student body and assigning players to roles that align with their unique skills and strengths. Traditional selection methods are often subjective and may overlook potential talent due to biases or lack of comprehensive evaluation. To address this, we propose a data-driven approach leveraging machine learning and statistical analysis to ensure fairness, objectivity, and optimal team performance. This project aims to develop predictive models using an extensive dataset of professional football players, analyzing attributes that contribute to player performance and suitability for specific roles. The methodology ensures that selections are based on quantifiable data, reducing biases and enhancing the team's overall competitiveness.

2 Methods

2.1 Data Collection

The study utilized a comprehensive dataset from FIFA, containing detailed attributes of professional football players. We connected to the SQLite database provided in Kaggle and loaded the *Player_Attributes* table into a pandas DataFrame. Initial inspection revealed that the dataset consisted of 183,978 observations and 42 columns, including both categorical and numerical variables. The dataset comprises various attributes that quantify different skills and abilities of football players.

2.2 Data Preprocessing

High-quality data is foundational for accurate modeling. The data cleaning process involved handling missing values, correcting incorrect entries, and addressing outliers. Irrelevant columns were dropped, and missing values were identified and treated. For categorical variables, the mode was imputed, while for numerical variables, mean or median imputation was used based on the distribution. Outliers were identified using z-scores and treated with capping (winsorization) to reduce skewness while retaining top-performing players' data.

2.3 Feature Engineering

Recognizing that different roles require specific skill sets, players were assigned to roles (Forward, Midfielder, Defender, and Goalkeeper) based on Role-specific scores. These scores were developed using weighted averages of relevant attributes, with weights determined based on the importance of each attribute to the role. Players were assigned to the role for which they had the highest score.

2.4 Machine Learning Approaches

We initially attempted to use the full dataset for modeling, but this approach resulted in excessively long run times. To address this, we decided to use PCA components for predictions instead. For unsupervised learning, we employed Principal Component Analysis (PCA) and K-Means Clustering. For supervised learning, we focused on predicting *overall_rating* using regression models such as Multiple Linear Regression, Ridge Regression, Lasso Regression, k-Nearest Neighbors (k-NN) Regressor, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regressor. Additionally, we employed stratified classification models to predict *player_role*, addressing the challenge of class imbalance in this feature.

3 Results

3.1 Unsupervised Learning

Principal Component Analysis (PCA) reduced the high-dimensional player attributes into two primary dimensions that aligned with football roles. The first dimension contrasted offensive contributions, such as crossing, dribbling, and finishing, with defensive attributes, including tackling and goalkeeping skills. The second dimension separated technical traits like agility and ball control from physical attributes such as strength and aggression. These dimensions provided a structured view of the data, naturally clustering players into four roles: attackers, midfielders, defenders, and goalkeepers. K-Means clustering, applied to both the PCA-transformed data and the full dataset, further validated these roles. While PCA facilitated visualization by capturing variance, clustering on the full dataset offered richer and more interpretable insights. The clusters corresponded to known football roles, reinforcing the validity of the engineered features.

3.2 Supervised Learning

We used a three-step method for model selection: first, we set aside a test sample for unbiased evaluation. Then, we performed hyperparameter tuning using cross-validation on training and validation subsets to identify the best model configuration. Finally, we compared different models through the same tuning process and evaluated the best performer on the test set. For *overall_ratings*, the k-NN Regressor achieved an adjusted R^2 of **0.9357** and (RMSE/y_{mean}) of **0.0261**. For *player_roles*, the kNN Classifier scored an F1 weighted average of **0.9**. Feature importance analysis led to the elimination of Component 4 in the k-NN model, resulting in a slight loss of performance for better interpretability and computational efficiency. Component 1 remained the most important feature. The optimized k-NN model achieved an adjusted R^2 of **0.9043** on the final test set.

4 Discussion

4.1 Practical Applications

The proposed data-driven approach eliminates biases in player selection, ensuring fairness and objectivity. This framework could revolutionize recruitment strategies in sports management by prioritizing skill over subjective assessments. By tailoring role assignments based on key attributes, the methodology enhances team performance and cohesive team dynamics. The framework's adaptability makes it a valuable tool for talent identification in youth academies and semi-professional clubs worldwide, improving scouting efficiency.

4.2 Business Relevance, Profitability, and Deployments

Football clubs and other team-based sports organizations can adopt this approach to build more effective teams. The detailed analysis of player attributes and their alignment with specific roles ensures that each player is utilized in a way that maximizes their strengths. This optimization can lead to better on-field performance and a more cohesive team dynamic. The methodology can be scaled to different levels of competition, from amateur to professional, making it a versatile tool for talent scouting. Additionally, this approach can be sold to teams training young talents and to more competitive teams looking to enhance their rosters. It can also be used to detect talented players in underperforming teams, providing a strategic advantage in player acquisition. Future work should focus on real-world validation by collecting data from Emlyon students and validating the framework in practical scenarios. Incorporating intangible factors such as teamwork, leadership, and adaptability, as well as addressing scalability challenges and limited external variables, will enhance the framework's practical utility. Continuous data collection and model refinement will further improve predictive capabilities, making the methodology more robust and applicable to various sports organizations.

5 Conclusion

The study presents a robust, data-driven methodology for player selection and role assignment in football, achieving fairness and performance optimization. By leveraging PCA, K-Means clustering, and supervised learning models, the approach ensures objective, accurate predictions while minimizing biases inherent in traditional selection methods. This methodology holds significant potential for Emlyon Business School and football clubs globally, revolutionizing talent scouting and team building in the sport. Future efforts should focus on real-world validation and the integration of intangible factors to further refine and expand the application of this approach.