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

CHAMAKHI Ayoub, CAUHAPE Agathe, PONCET Maude, PURKAYASTHA Aakash

emlyon business school

Professor Dr. Franck JAOTOMBO

December 4, 2024

Abstract

This study presents a data-driven approach to player selection and role assignment for Emlyon Business School's football team, leveraging machine learning and statistical analysis to optimize team performance and eliminate biases inherent in traditional selection methods. Using a comprehensive dataset of professional football players, the research employs Principal Component Analysis (PCA) for dimensionality reduction and K-Means clustering for role assignment. Supervised learning models, including k-Nearest Neighbors Regressor (k-NN) and k-Nearest Neighbors Classifier (k-NN), predict overall ratings and player roles with high accuracy. The methodology enhances objectivity in player recruitment and role allocation, ensuring that players are assigned to positions based on their strengths and skills. The proposed approach has broad applications in sports management, including talent identification in youth academies and semi-professional clubs, and offers potential for improving team performance across various levels of competition. Future work will focus on real-world validation and incorporating intangible factors like teamwork and adaptability.

Keywords: *Football Analytics, Players Rating, Player Roles, Machine Learning, PCA, K-Means, kNN Regressor, kNN Classifier*

Link to Notebook and Dataset: Google Drive

*Group 1 project for the course Principles & Elementary Models, under the supervision of Professor Dr. Franck JAOTOMBO.

Contents

1 Introduction

Emlyon Business School is dedicated to bringing together the most talented players to build a strong and dynamic soccer club. The endeavor requires not only identifying the best football talent within the student body but also assigning players to roles that align with their unique skills and strengths. Traditional selection methods may be subjective and overlook potential talent due to biases or lack of comprehensive evaluation. To address this challenge, we propose a data-driven approach that leverages machine learning and statistical analysis to ensure fairness, objectivity, and optimal team performance.

Our approach involves developing predictive models using an extensive dataset of professional football players. By analyzing attributes that contribute to player performance and suitability for specific roles, we aim to create a framework that can be applied to the student population. This methodology ensures that selections are based on quantifiable data, reducing biases and enhancing the team's overall competitiveness.

This paper details the process from data collection and cleaning to the application of advanced analytical techniques. We explore both supervised and unsupervised learning models, interpret the outcomes as data scientists, and provide insights that can guide Emlyon Business School in assembling a formidable football team. Additionally, this approach holds the potential to be scaled and utilized by football clubs worldwide for talent identification among young players.

2 Data Collection

To develop robust predictive models, we utilized a comprehensive Kaggle dataset containing detailed attributes of FIFA's professional football players. This

dataset, comprising 183,978 observations and 42 columns, served as a valuable proxy for modeling and predicting the performance of student players at Emlyon Business School. It included a wide array of attributes such as technical skills, physical abilities, and goalkeeping skills.

We connected to the SQLite database provided in the dataset and loaded the *Player_Attributes* table into a pandas DataFrame. An initial inspection of the data revealed its structure and content, highlighting both categorical and numerical variables. This step was critical for guiding the subsequent cleaning and analysis processes.

3 Data Overview

The dataset comprises various attributes that quantify different skills and abilities of football players. Each variable represents a specific aspect of a player's performance and is scored on a standardized scale, typically ranging from 1 to 100. Below is a brief description of key variables and the meaning of their scoring:

3.1 Variable Descriptions

- **Potential**: Indicates the player's potential to develop and improve over time. A higher score reflects a greater capacity for growth.
- **Attacking Work Rate** and **Defensive Work Rate**: Categorical variables (*low*, *medium*, *high*) that describe the player's tendency to participate in offensive and defensive plays.
- **Crossing**: Measures the player's ability to deliver accurate crosses into the penalty area. Higher scores indicate better crossing skills.

- **Finishing**: Reflects the player's proficiency in converting goal-scoring opportunities. A higher score denotes better finishing ability.
- **Heading Accuracy**: Assesses how accurately a player can head the ball during offensive and defensive plays.
- **Short Passing** and **Long Passing**: Evaluate the accuracy and effectiveness of a player's passes over short and long distances.
- **Dribbling** and **Ball Control**: Indicate the player's ability to maneuver the ball while evading opponents and maintain possession.
- **Acceleration** and **Sprint Speed**: Reflect the player's speed attributes, with higher scores indicating faster acceleration and top speed.
- **Agility** and **Balance**: Assess the player's ability to change direction quickly and maintain stability.
- **Strength** and **Jumping**: Indicate the player's physical power and ability to leap vertically.
- **Aggression** and **Interceptions**: Measure the player's intensity and ability to read and intercept the game.
- **Goalkeeping Attributes**: Specialized attributes for goalkeepers, including *gk_diving*, *gk_handling*, *gk_kicking*, *gk_positioning*, and *gk_reflexes*.
- **Overall Rating**: An aggregated score representing the player's overall ability, often used as a benchmark for performance.

These scores are typically on a scale from 1 to 100, where higher values represent better performance in that attribute. Understanding these variables is essential for modeling player performance and role suitability.

4 Data Cleaning

High-quality data is foundational for accurate modeling. Our data cleaning process focused on handling missing values, correcting incorrect entries, and addressing outliers to ensure the dataset's integrity.

4.1 Dropping Irrelevant Variables

We identified and removed irrelevant columns such as *date*, *player_fifa_api_id*, *player_api_id*, and *id*. These variables were not directly related to player performance or attributes and could introduce unnecessary noise into our models. By focusing on pertinent variables, we streamlined the dataset for more effective analysis.

4.2 Handling Missing Values

4.2.1 Identifying Missing Values

We began by listing and plotting missing values across categorical and numerical columns. The initial assessment revealed that certain columns had missing values, with *attacking_work_rate* and several numerical attributes like *volleys*, *curve*, *agility*, *balance*, *jumping*, *vision*, and *sliding_tackle* having the highest counts.

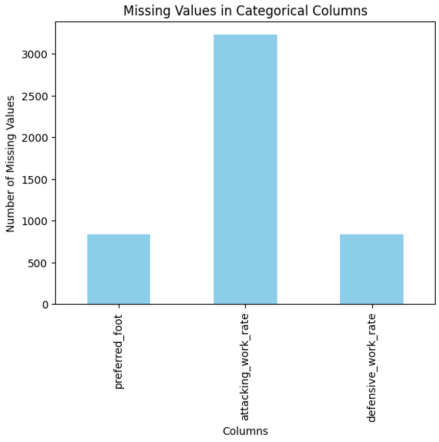

Figure 1: Missing Categorical Data

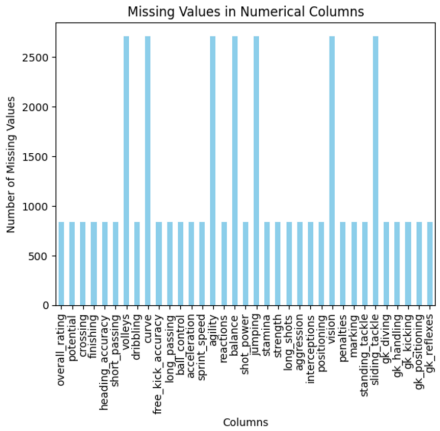

Figure 2: Missing Numerical Data

Figure 3: Comparison of Missing Data in Categorical and Numerical Variables

4.2.2 Dropping Rows with Missing Values

Given the dataset's size of 183,978 rows, dropping rows with 836 missing values would result in a minimal reduction of approximately 0.45%. This minimal reduction in size supported our decision to drop these rows, as it would not significantly impact the dataset's representativeness or the quality of our analysis.

4.2.3 Imputing Missing Values

After dropping rows with missing values in key columns, we were left with missing values in both categorical and numerical columns.

For the remaining missing categorical values in *attacking_work_rate*, we imputed the mode (the most frequent value) to fill in the missing entries. This approach maintained the distribution of categories in the dataset.

For numerical columns with missing values, such as *volleys*, *curve*, *agility*, *balance*, *jumping*, *vision*, and *sliding_tackle*, we determined the appropriate imputation method by comparing the mean and median values. We calculated the

relative difference between the mean and median for each attribute:

- If the relative difference was greater than 10%, indicating skewness due to outliers, we used median imputation.
- If the relative difference was less than or equal to 10%, indicating a relatively symmetric distribution, we used mean imputation.

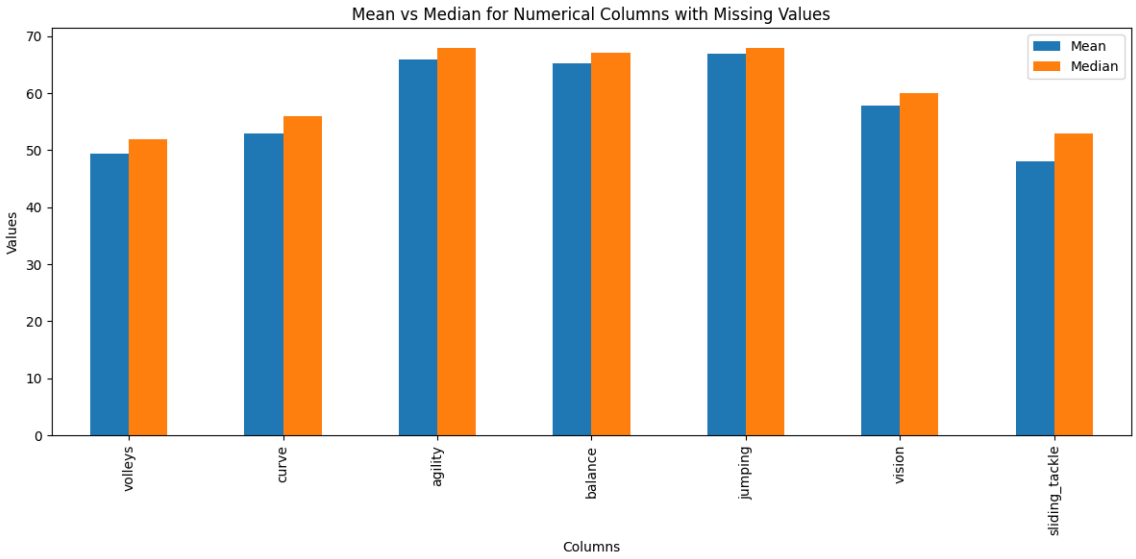

Figure 4: Imputing Missing Values

This method ensured that the imputed values reflected the underlying distribution of each attribute, minimizing the impact of outliers.

4.3 Correcting Incorrect Categorical Entries

We addressed incorrect entries in categorical variables by replacing invalid values with the mode. Specifically, we ensured that *attacking_work_rate* and *defensive_work_rate* contained only valid categories (*low*, *medium*, *high*). This correction was essential for maintaining consistency and accuracy in categorical data, which could significantly affect classification models.

4.4 Identifying and Treating Outliers

Outliers can distort statistical analyses and machine learning models. We identified outliers in numerical columns using z-scores, considering values with an absolute z-score greater than 3 as outliers. Visualization through histograms and box plots revealed that some attributes had significant outliers, particularly at the higher end of the scale.

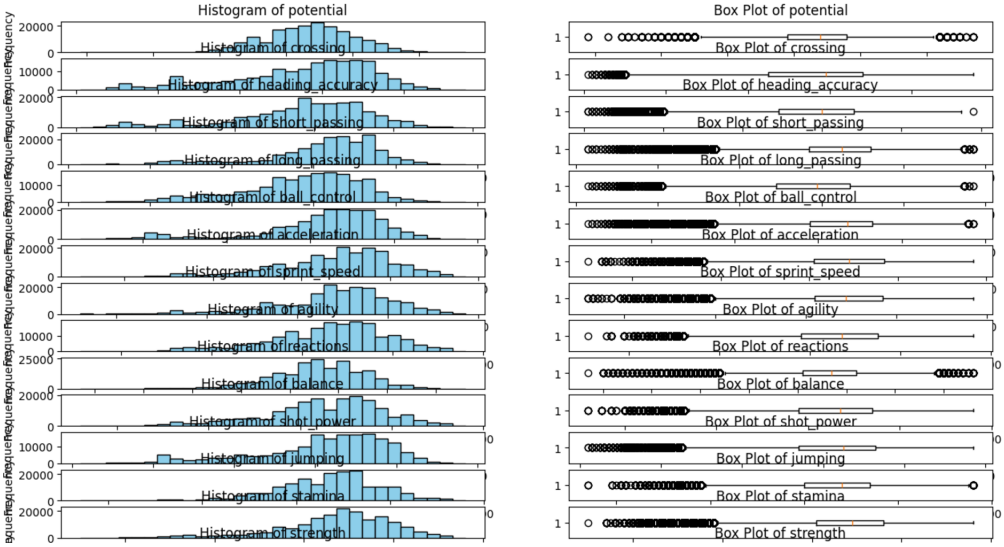

Figure 5: Outliers Visualizations

To handle outliers without discarding data, we applied capping (winsorization) to certain attributes. For high-performance attributes where outliers represented exceptionally skilled players (e.g., *sprint_speed*, *acceleration*), we capped the outliers at the 99th percentile, retaining top-performing players' data while reducing skewness. For attributes where outliers might indicate data entry errors (e.g., *balance*, *penalties*), we capped values at the 95th percentile to correct anomalies.

5 Analyzing Data

With a clean dataset, we proceeded to analyze the data to uncover patterns and relationships among player attributes. This step was crucial for feature selection and understanding the underlying structure of the data.

5.1 Feature Engineering

Recognizing that different roles require specific skill sets, we assigned players to roles (*Forward*, *Midfielder*, *Defender*, and *Goalkeeper*) based on key attributes. We developed role-specific scores using weighted averages of relevant attributes. The weights were determined based on the importance of each attribute to the role, informed by domain knowledge.

5.1.1 Weighting of Attributes

- **Forward Score**:
  - *Sprint Speed*: 30%
  - *Dribbling*: 30%
  - *Shot Power*: 20%
  - *Agility*: 20%
- **Midfielder Score**:
  - *Short Passing*: 30%
  - *Long Passing*: 30%
  - *Vision*: 20%
  - *Ball Control*: 20%
- **Defender Score**:

  - *Strength*: 30%
  - *Jumping*: 30%
  - *Aggression*: 20%
  - *Heading Accuracy*: 20%

- **Goalkeeper Score**:
  - *Goalkeeper Reflexes*: 30%
  - *Goalkeeper Diving*: 30%
  - *Goalkeeper Handling*: 20%
  - *Goalkeeper Kicking*: 20%

These weights reflect the relative importance of each attribute to the specific role. For example, *Sprint Speed* and *Dribbling* are crucial for a forward's ability to outpace defenders and navigate through tight spaces, hence they carry higher weights.

Players were assigned to the role for which they had the highest score. This stratification allowed us to tailor our analysis and modeling to the nuances of each role, recognizing that the importance of certain attributes varies by position.

5.2 Descriptive Statistics by Role

We computed descriptive statistics for each role to validate our stratification. The analysis revealed that:

- *Forwards* had higher mean values in attributes crucial for offensive play, such as *finishing*, *dribbling*, and *acceleration*.
- *Midfielders* excelled in *passing*, *vision*, and *stamina*, reflecting their role as the team's link between defense and attack.

- *Defenders* showed elevated means in *tackling*, *strength*, and *aggression*, aligning with the physical demands of defensive positions.
- *Goalkeepers* specialized in goalkeeping attributes, with high scores in *GK Reflexes*, *GK Diving*, and *GK Handling*.

These findings confirmed that our role-based stratification accurately reflected the distinct attribute profiles expected in professional football.

5.3 Exploratory Data Analysis

We performed univariate and bivariate analyses to gain deeper insights.

5.3.1 Univariate Categorical Data

The dataset of football players includes several univariate categorical variables that provide valuable insights for predicting *overall_rating* and *player_role*. The majority of players (75.57%) prefer their right foot, which can influence their playing style and effectiveness in certain roles, potentially affecting overall ratings. Most players (71.94%) have a medium attacking work rate, indicating a balanced approach to attacking. High work rates (23.38%) might be crucial for offensive roles, while low rates (4.68%) could suggest more defensive or supportive roles. Similarly, most players (75.17%) have a medium defensive work rate, with high defensive work rates (14.77%) being important for defenders and low rates (10.06%) possibly more common among forwards. The distribution of player roles shows a notable imbalance, with forwards (39.96%) and defenders (36.97%) being the most common, followed by midfielders (15.00%) and goalkeepers (8.06%). This imbalance needs to be addressed later with stratification techniques to ensure that the predictive model is not biased towards the more frequent roles. Understanding these distributions and characteristics can help

tailor training programs and strategies to enhance player performance and overall team effectiveness. For instance, focusing on improving the defensive work rate of midfielders or the attacking work rate of forwards could lead to better overall ratings and more specialized roles.

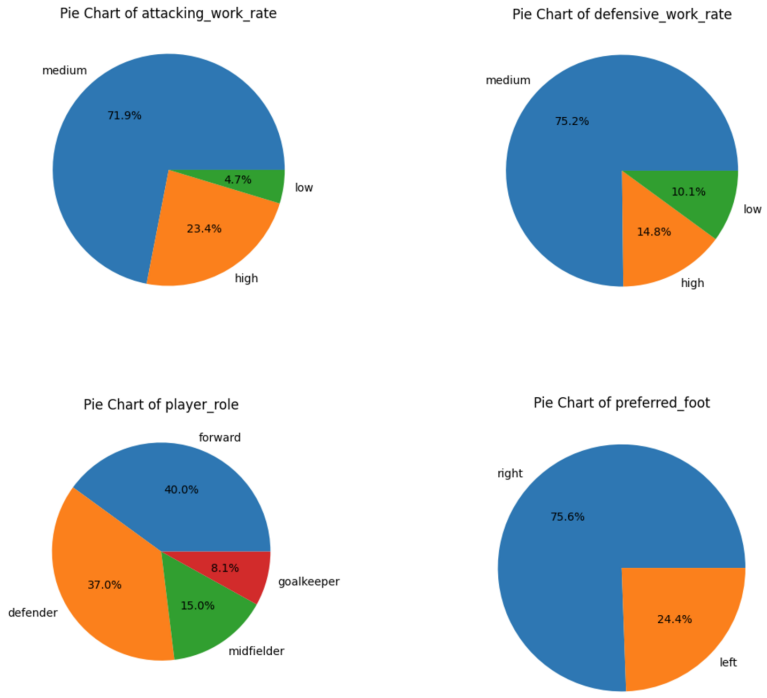

Figure 6: Univariate categorical variables

5.3.2 Univariate Numerical Variables

The histograms provide an overview of the distribution for various numerical attributes. Most variables, such as *overall_rating*, potential, and dribbling, show a roughly normal distribution centered between 50 and 80, indicating the majority of players perform within this range. Attributes like crossing, *heading_accuracy*,

and *long_shots* exhibit slight right skewness, suggesting fewer players excel at high values. Goalkeeping attributes, including *gk_diving* and *gk_handling*, have distinct distributions, as they are specific to goalkeepers and feature concentration near the lower end for non-goalkeepers. This differentiation highlights the positional specialization in player attributes. For correlation analysis, Pearson correlation can be used to assess linear relationships, as many variables display normal distributions. However, caution should be taken with skewed variables like crossing and *heading_accuracy*, where transformations may be needed to ensure accurate results.

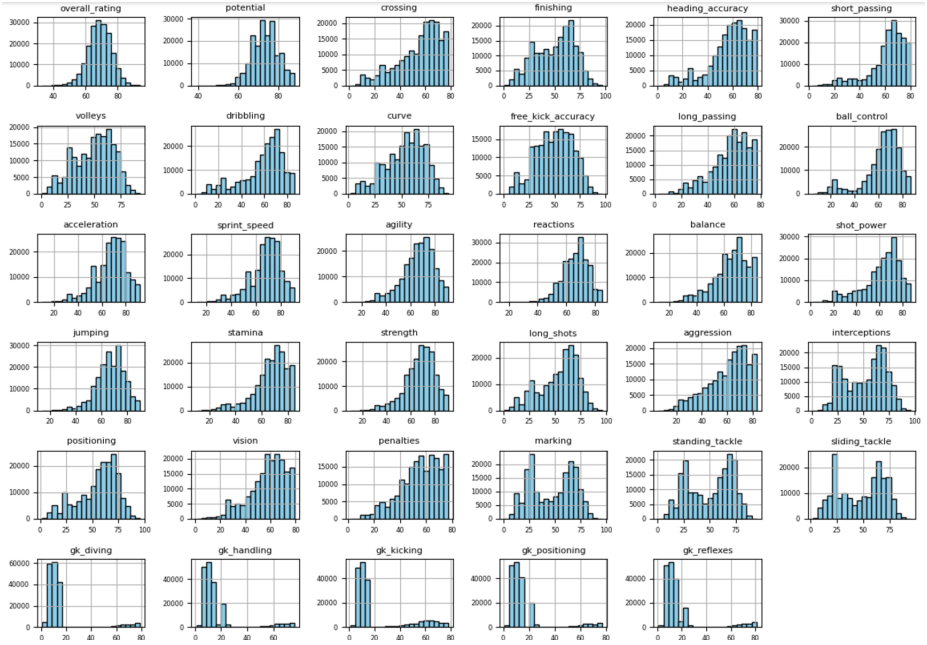

Figure 7: Univariate Numerical Variables

5.3.3 Bivariate Numerical Variables

The analysis of the dataset reveals several key insights that can be leveraged for predicting overall player ratings and classifying player roles in football. The univariate analysis of numerical data shows that most player attributes, such as

crossing, finishing, and dribbling, follow a right-skewed distribution, indicating that a majority of players have lower to mid-range skill levels, with fewer players achieving high skill levels. This pattern is consistent across technical, physical, and specialized attributes, suggesting that top-tier skills are rare. The correlation matrix highlights strong positive correlations between overall rating and potential ($r = 0.91$), as well as between short passing and ball control ($r = 0.91$), indicating that players with high overall ratings tend to have high potential and that short passing skills are closely linked to ball control. Moderate positive correlations, such as those between finishing and shot power ($r = 0.70$), and dribbling and agility ($r = 0.70$), suggest that these attributes often co-occur in skilled players.

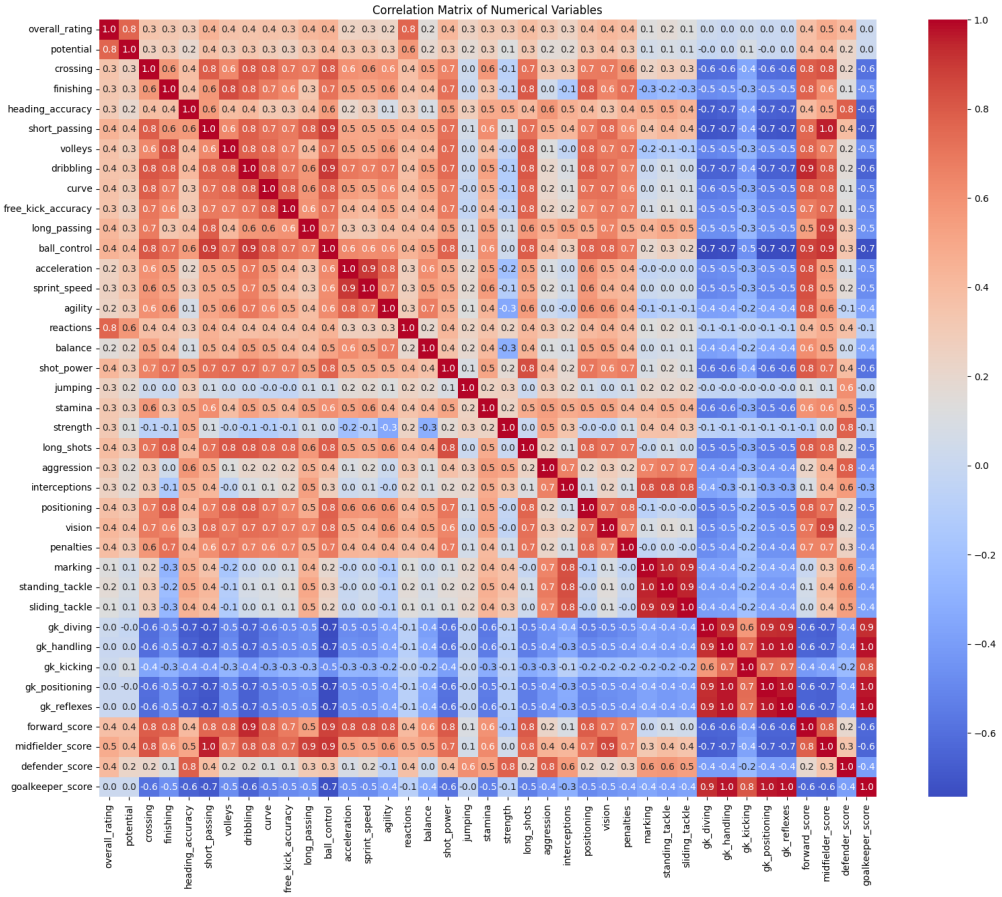

Figure 8: Correlation Matrix

The analysis of mixed categorical and quantitative data reveals significant differences in various attributes based on preferred foot, attacking work rate, defensive work rate, and player role. For instance, left-footed players tend to have better crossing ($F = 5675.72$, $p < 0.001$), dribbling ($F = 1320.94$, $p < 0.001$), and curve abilities ($F = 3068.10$, $p < 0.001$) compared to right-footed players. Players with high attacking work rates generally have better overall ratings ($F = 2367.19$, $p < 0.001$), potential ($F = 1959.20$, $p < 0.001$), and technical skills like crossing ($F = 8398.06$, $p < 0.001$), finishing ($F = 9077.99$,

$p < 0.001$), and dribbling ($F = 11174.84$, $p < 0.001$). Conversely, players with high defensive work rates excel in attributes like heading accuracy ($F = 3759.57$, $p < 0.001$), interceptions ($F = 11832.88$, $p < 0.001$), and tackling ($F = 14532.03$, $p < 0.001$). Player roles show distinct patterns, with midfielders having the highest overall ratings ($F = 1011.16$, $p < 0.001$) and potential ($F = 1756.04$, $p < 0.001$), defenders excelling in strength ($F = 23084.59$, $p < 0.001$) and aggression ($F = 34379.63$, $p < 0.001$), forwards in finishing ($F = 42627.57$, $p < 0.001$) and dribbling ($F = 114837.36$, $p < 0.001$), and goalkeepers in specialized goalkeeping skills ($F > 400000$, $p < 0.001$).

The chi-square tests indicate significant associations between categorical variables, such as preferred foot and attacking work rate ($\chi^2 = 358.498$, $p < 0.001$, $V = 0.044$), preferred foot and defensive work rate ($\chi^2 = 294.087$, $p < 0.001$, $V = 0.040$), and player roles with work rates ($\chi^2 > 16000$, $p < 0.001$, $V > 0.20$). These associations suggest that certain attributes are more prevalent in specific player roles and work rates, which can be crucial for classification tasks. For regression of overall rating, attributes like potential, short passing, ball control, and acceleration, which show strong correlations, can be valuable predictors. For classifying player roles, the distinct patterns in attributes like crossing, finishing, heading accuracy, and goalkeeping skills can be utilized to differentiate between defenders, forwards, midfielders, and goalkeepers. Overall, these insights can enhance the accuracy of predictive models for player ratings and role classification in football.

6 Data Preparation for Modeling

Before modeling, we encoded categorical variables into numerical formats suitable for machine learning algorithms. Ordinal variables like *attacking_work_rate* and *defensive_work_rate* were mapped to numerical values (*low*: 0, *medium*: 1,

high: 2). Binary variables like *preferred_foot* were one-hot encoded, creating a binary variable *preferred_foot_right*.

Standardization was applied to numerical features using z-scores to ensure that variables were on the same scale, preventing features with larger scales from disproportionately influencing the models. This preprocessing step was critical for algorithms sensitive to feature scaling, such as k-Nearest Neighbors and gradient descent-based methods.

7 Machine Learning

With the data prepared, we applied both unsupervised and supervised machine learning techniques to extract insights and develop predictive models.

7.1 Unsupervised Learning

7.1.1 Principal Component Analysis

Before applying Principal Component Analysis (PCA), it is essential to verify that the data meets the necessary prerequisites. One critical assumption for PCA is that the variables should have some level of correlation among them. To assess this, we conducted Bartlett's test of Sphericity. This test evaluates the hypothesis that the correlation matrix of the variables is an identity matrix, implying no correlation among variables. The test yielded a Bartlett's statistic of approximately 8358759.179 with a p-value of 0.000. The extremely low p-value indicates significant differences in variances, suggesting that the variables are indeed correlated, and thus, the data is suitable for PCA. Another crucial measure to assess the adequacy of the data for PCA is the Kaiser-Meyer-Olkin (KMO) test. The KMO test evaluates the sampling adequacy for each variable and the overall model. The KMO score ranges from 0 to 1, with values closer

to 1 indicating that the data is highly suitable for factor analysis. Our dataset yielded a KMO model score of 0.96, which is exceptionally high and confirms that the data is highly suitable for PCA. This high KMO score reinforces the appropriateness of applying PCA to our football dataset. Eigenvalues represent the amount of variance captured by each principal component. By examining the eigenvalues, we can determine the number of components that capture most of the variance in the data. The first few eigenvalues were significantly larger, indicating that these components capture the majority of the variance. This insight is crucial for deciding the number of principal components to retain in the analysis. The scree plot is a visual tool that helps in determining the optimal number of principal components to retain. The plot displays the eigenvalues against the number of components. The elbow method, which looks for a point where the eigenvalues start to level off, suggested that 3 factors might be sufficient. However, applying the lambda = 1 rule, which retains components with eigenvalues greater than 1, indicated that 6 factors should be chosen. This discrepancy highlights the importance of considering multiple criteria when deciding the number of components to retain.

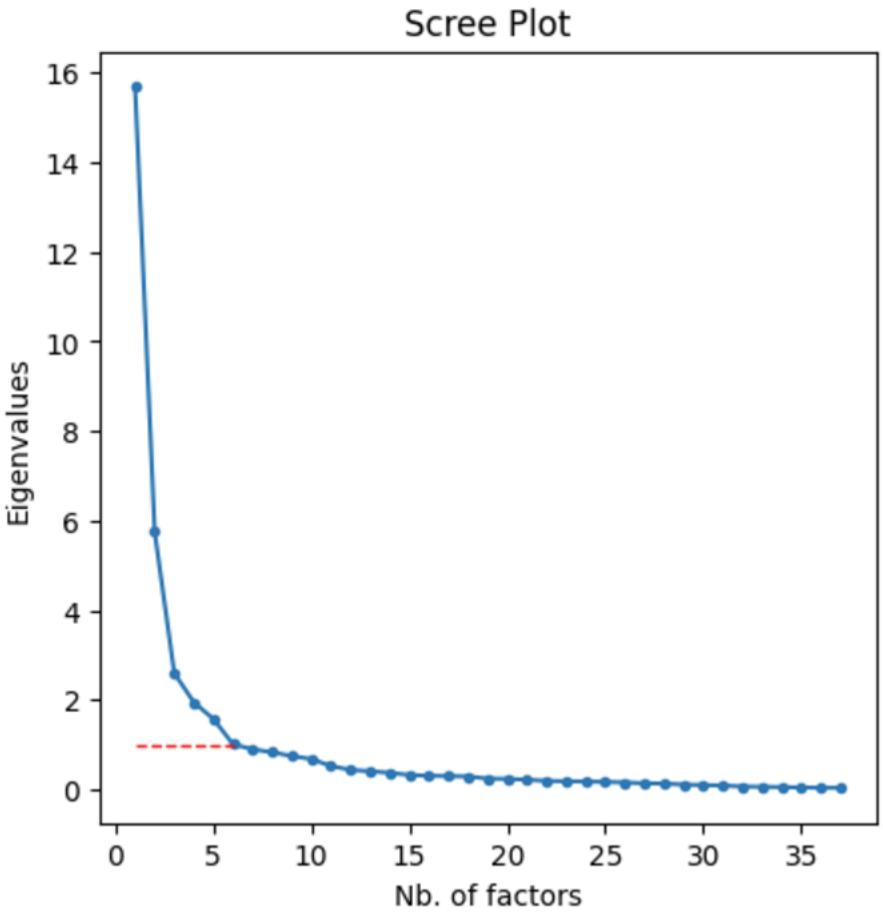

Figure 9: PCA Screeplot

The explained variance plot shows the cumulative percentage of variance explained by the principal components. This plot is essential for understanding how many components are needed to explain a significant portion of the variance in the data. With 2 factors, only 60% of the variance is explained, while with 5 factors, we can explain 80%. Adding further factors results in a steep increase in explained variance. This insight is crucial for balancing the trade-off between the number of components and the variance explained.

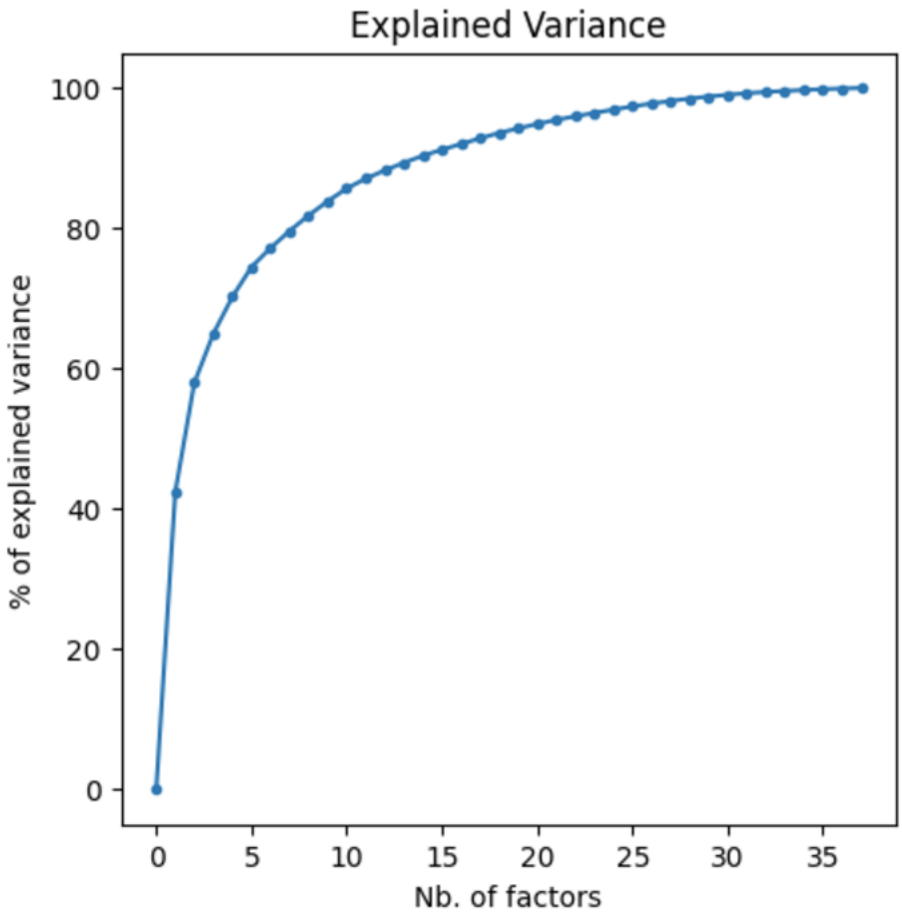

Figure 10: PCA % of Explained Variance

The Karlis-Saporta-Spinaki threshold is another method to determine the number of components to retain. This threshold is calculated based on the number of variables and observations. The threshold value was approximately 1.028, further supporting the choice of selecting 5 to 6 factors. This threshold provides additional validation for the number of components to retain and ensures that the chosen components capture a significant portion of the variance.

The broken sticks method is a graphical approach to determine the number of components to retain. This method involves plotting the eigenvalues and

comparing them to a threshold derived from the broken sticks. The broken sticks are leaning towards a 2-factor solution. However, considering our knowledge of football and the number of roles a footballer can have, we proceeded with choosing 4 factors. This decision aligns with the practical aspects of football and ensures that the components are interpretable and meaningful.

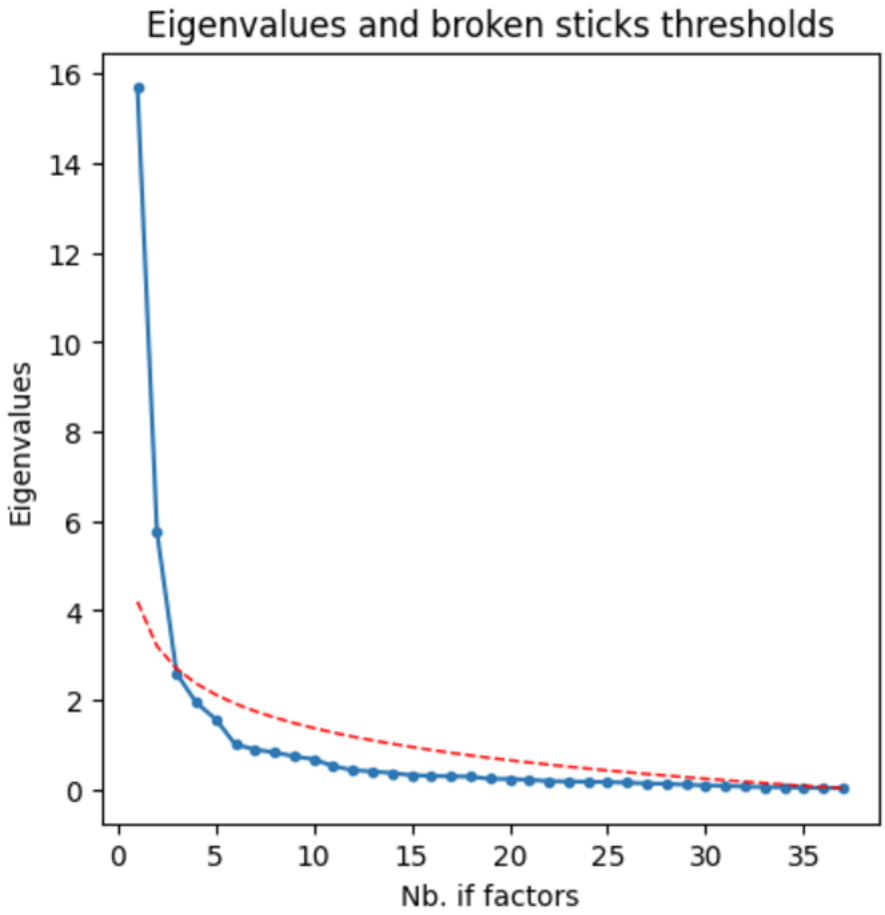

Figure 11: PCA Brokensticks

The correlation circle is a graphical representation that shows the correlation between the original variables and the principal components. By examining the correlation circle, we can identify which variables contribute most to each

principal component.

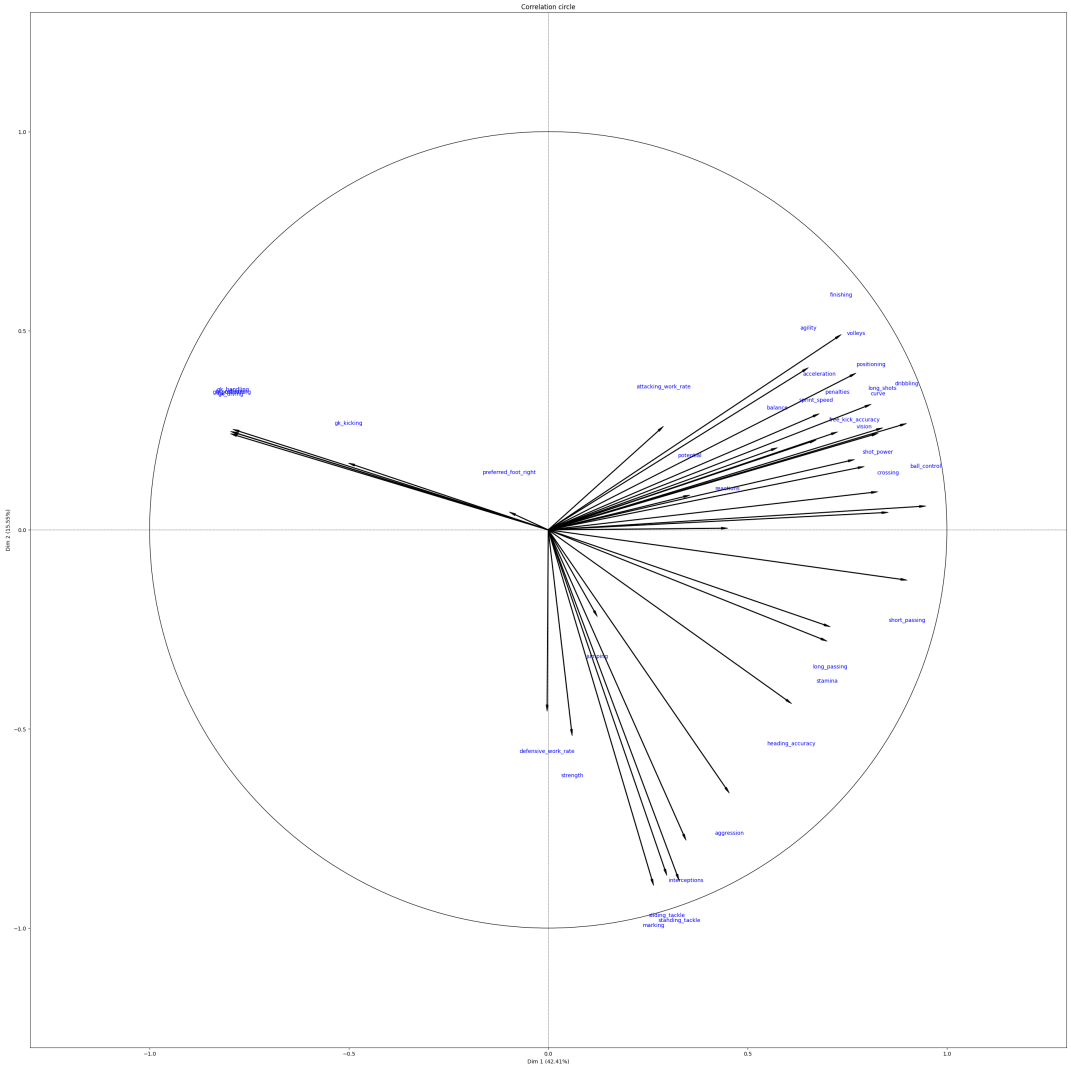

Figure 12: PCA Correlation Circle

The correlation circle derived from the Principal Component Analysis (PCA) on the football dataset reveals two primary dimensions and four distinct clusters that correspond to different player roles. The first dimension, represented by the horizontal axis, differentiates between offensive and defensive contributions.

Attributes such as crossing, short passing, dribbling, ball control, and finishing, which are crucial for midfielders and forwards, are positively correlated and found on the right side, indicating their role in creating and converting scoring opportunities. Conversely, goalkeeping skills like GK diving, GK handling, GK positioning, and GK reflexes are negatively correlated and positioned on the left side, highlighting their unique defensive role. The second dimension, represented by the vertical axis, distinguishes between technical and physical contributions. Attributes like finishing, agility, acceleration, sprint speed, dribbling, and curve, essential for forwards and attacking midfielders, are positively correlated and located at the top, emphasizing their importance in scoring and creating chances. Meanwhile, defensive skills such as marking, standing tackle, sliding tackle, interceptions, strength, and aggression are negatively correlated and found at the bottom, underscoring their significance in preventing the opposition from scoring.

The correlation circle can be divided into four clusters: the top right quadrant, which includes forwards and attacking midfielders characterized by finishing, agility, dribbling, curve, and free kick accuracy; the bottom right quadrant, representing midfielders with strong passing, ball control, and vision; the bottom left quadrant, encompassing defenders with robust tackling, marking, and interception skills; and the top left quadrant, featuring goalkeepers with specialized skills like diving and handling. These dimensions and clusters provide a comprehensive understanding of the key attributes defining different player roles in football, which can be leveraged for further analysis and predictive modeling.

After selecting four components for Principal Component Analysis (PCA) and retraining the model, the resulting components were identified as follows: F1: Forward, representing features associated with attacking players and goal-scoring attributes; F2: Defenders, capturing characteristics linked to defensive

actions and positioning; F3: Goalkeepers, highlighting attributes unique to goalkeeping performance; and F4: Midfielders, encompassing traits indicative of players who contribute both defensively and offensively in central roles.

7.1.2 K-Means Cluster Analysis

The K-Means clustering analysis was conducted using both the PCA-transformed dataset and the full dataset to identify distinct player roles based on their attributes. The elbow method was employed to determine the optimal number of clusters by plotting the inertia (sum of squared distances from each point to its assigned cluster center) against the number of clusters. For both the PCA-transformed dataset and the full dataset, the inertia plots showed a significant decrease up to approximately $k = 4$ clusters, suggesting that 4 clusters provided a good balance between model complexity and explanatory power.

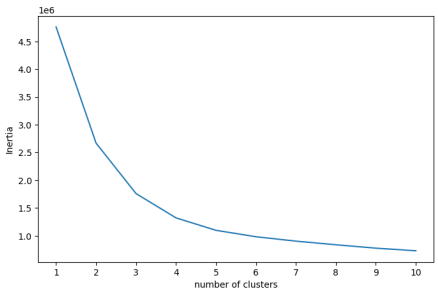

Figure 13: KMeans Clustering PCA Inertia

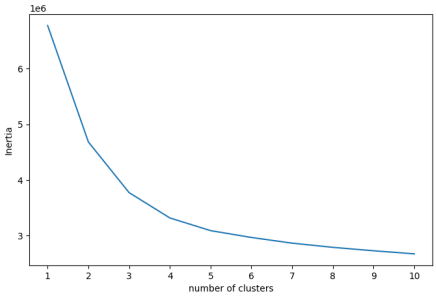

Figure 14: Kmeans Clustering Full Dataset Inertia

Figure 15: KMeans Clustering Inertias

Initially, K-Means clustering was performed on the PCA-transformed dataset, focusing on the first four principal components. While the elbow plot confirmed that 4 clusters were optimal, the cluster centers for the PCA components did not yield clear or interpretable groupings.

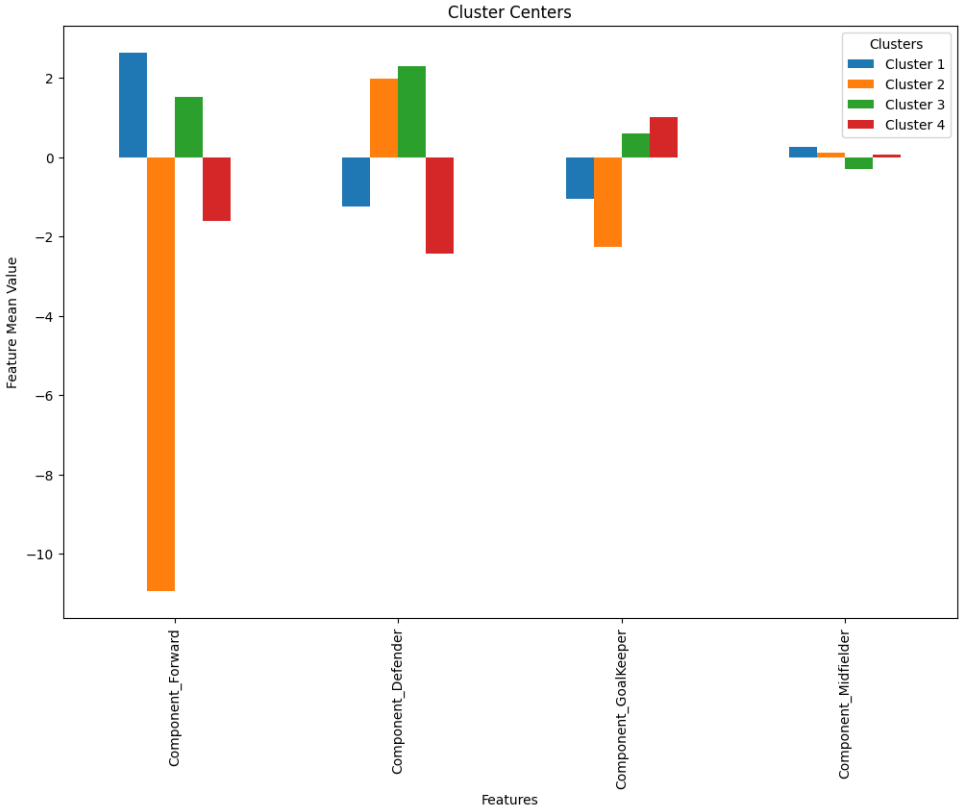

Figure 16: KMeans Clustering PCA

This may be attributed to PCA maximizing variance but not necessarily preserving the inherent clustering structure of the original data. The principal components capture the directions of maximum variance, but these directions may not align with clusters defined by player roles.

Subsequently, K-Means clustering was applied to the full dataset, including all original features. Clustering with $k = 4$ produced meaningful and interpretable clusters corresponding to distinct player roles.

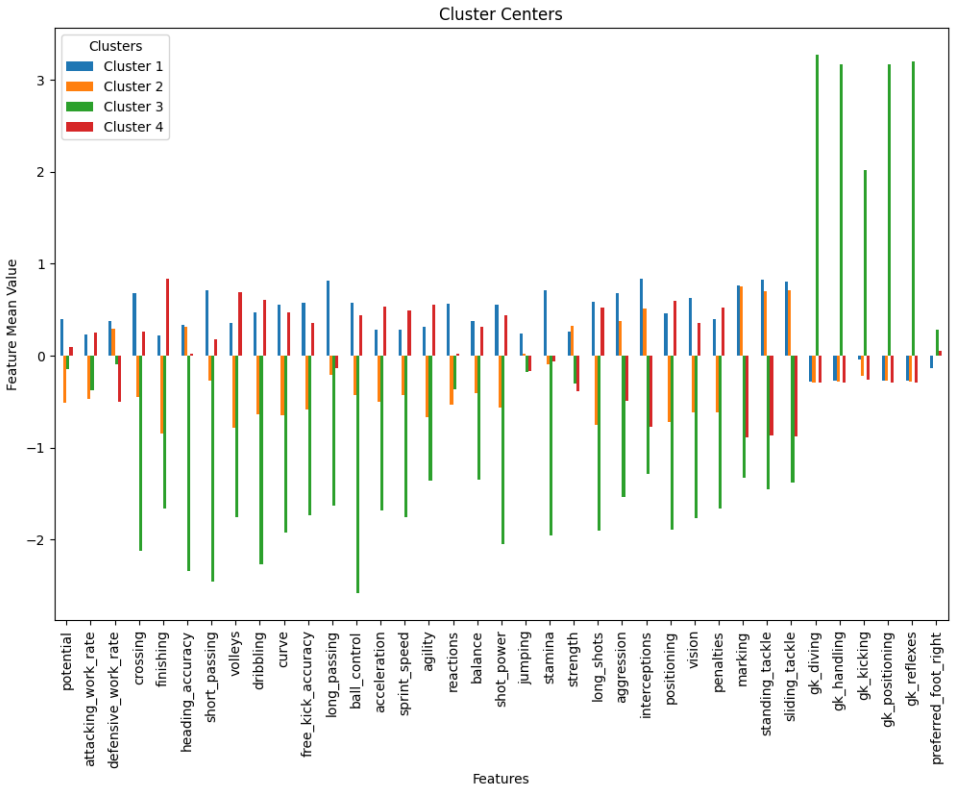

Figure 17: KMeans Clustering PCA

The cluster centers revealed key insights into the characteristics of each role:

- **Cluster 1 (Midfielders)**: Players with balanced attributes, excelling in passing, dribbling, and stamina. Midfielders require versatility to contribute to both attacking and defensive play, controlling the game and creating opportunities.
- **Cluster 2 (Goalkeepers)**: Players with high values in goalkeeper-specific attributes such as GK diving, GK handling, GK kicking, GK positioning, and GK reflexes. These players have specialized goalkeeping skills, while other attributes remain average or below average, typical for their focused role.

- **Cluster 3 (Defenders)**: Players strong in defensive skills like standing tackle, sliding tackle, marking, interceptions, and strength. These players excel in preventing opposition plays through physical robustness and defensive expertise.

- **Cluster 4 (Attackers)**: Players with high values in attacking attributes such as finishing, shot power, dribbling, acceleration, and sprint speed. These players specialize in scoring goals and creating offensive opportunities.

The K-Means clustering analysis on the full dataset provided clearer and more interpretable clusters than the PCA-transformed data clustering. The resulting clusters effectively corresponded to distinct player roles: midfielders, goalkeepers, defenders, and attackers. This approach confirmed that certain combinations of attributes naturally group players into specific roles, offering valuable insights for player evaluation and team composition.

8 Supervised Learning

Upon trying the full dataset on supervised learning, we discovered that it was too large, and models outside of basic regression models took an excessive amount of time to render. To address this, we decided to use PCA, reducing the dataset to 4 components derived from unsupervised analysis. Furthermore, to tackle the dataset effectively, we employed an advanced method for model selection. This method consists of three main parts. First, we selected a test sample, which was set aside and never used during the subsequent phases to ensure unbiased evaluation. Second, we performed hyperparameter tuning by defining a grid of potential hyperparameter values. The model's main training set was divided into training and validation subsets. Using cross-validation, the model

was trained and validated multiple times for each hyperparameter combination. The performance scores were averaged, and the model configuration yielding the best score was saved. Finally, we compared different tuned models by subjecting multiple learner models to the same hyperparameter tuning process. The best-performing model was identified and evaluated on the previously untouched test set to compute the test score, ensuring the model's generalization ability.

8.1 Linear Outcome: Predicting `overall_rating`

Our first target variable, `overall_rating`, is continuous, making it suitable for regression analysis. We split the dataset into training and test sets, ensuring robust evaluation of our models.

Multiple Linear Regression model demonstrated strong performance with an average adjusted R^2 of 0.6874 and a normalized RMSE of 0.0574 during cross-validation. On the test set, the model achieved an adjusted R^2 of 0.6920 and a normalized RMSE of 0.0569. The final test set results were similarly robust, with an adjusted R^2 of 0.6960 and a normalized RMSE of 0.0567. This model works well for linear relationships and provides a baseline for more complex models. It is simple and interpretable, making it a good starting point.

Ridge Regression adds L2 regularization to handle multicollinearity and overfitting, improving the model's generalization ability. The best model, with α set to 1, showed excellent performance with a cross-validation RMSE of 3.9357. On the test set, the model achieved an adjusted R^2 of 0.6920 and a normalized RMSE of 0.0569. The final test set results were consistent, with an adjusted R^2 of 0.6960 and a normalized RMSE of 0.0567.

Lasso Regression adds L1 regularization, which can shrink some coefficients to zero, effectively performing feature selection and improving model interpretability. The best α value was found to be 0.01, yielding a cross-validation RMSE of 3.9357. The test set performance was strong, with an adjusted R^2 of 0.6920 and a normalized RMSE of 0.0569. The final test set results mirrored these findings, with an adjusted R^2 of 0.6960 and a normalized RMSE of 0.0567.

k-Nearest Neighbors (k-NN) Regressor was evaluated with different values of k. The best performance was achieved with $k = 3$, resulting in a cross-validation RMSE of 1.8537. The test set performance was exceptional, with an adjusted R^2 of 0.9336 and a normalized RMSE of 0.0264. The final test set results were equally impressive, with an adjusted R^2 of 0.9357 and a normalized RMSE of 0.0261. KNN is a non-parametric method that captures complex relationships without assuming linearity. It performs well with sufficient data and is robust to outliers.

Decision Tree Regressor was tuned using grid search with 10-fold cross-validation. The best model had a *max depth* of 15, *min samples split* of 5, and *min samples leaf* of 5. This model achieved a cross-validation RMSE of 2.0691. On the test set, the model performed well with an adjusted R^2 of 0.9145 and a normalized RMSE of 0.0300. The final test set results were similarly strong, with an adjusted R^2 of 0.9163 and a normalized RMSE of 0.0297. Decision Trees can capture non-linear relationships and interactions between features. They are interpretable and can handle both numerical and categorical data.

Random Forest Regressor was tuned with a simplified hyperparameter grid. The best model had 100 *estimators* and a *max depth* of 10. This model achieved a cross-validation RMSE of 2.3188. On the test set, the model performed well with an adjusted R^2 of 0.8929 and a normalized RMSE of 0.0336.

The final test set results were consistent, with an adjusted R^2 of 0.8934 and a normalized RMSE of 0.0336. Random Forests combine multiple decision trees to reduce overfitting and improve generalization. They are robust and perform well with large datasets.

XGBoost Regressor Finally, we evaluated an XGBoost Regressor with hyperparameter tuning for learning rate and number of estimators. The best model had a *learning rate* of 0.2 and 200 *estimators*. This model achieved a cross-validation RMSE of 2.2491. On the test set, the model performed well with an adjusted R^2 of 0.9004 and a normalized RMSE of 0.0324. The final test set results were similarly strong, with an adjusted R^2 of 0.8997 and a normalized RMSE of 0.0326. XGBoost is an efficient and scalable implementation of gradient boosting, which builds an ensemble of decision trees sequentially. It handles missing values, performs well with imbalanced data, and is highly customizable.

Conclusion for Regression Among the regression models evaluated, the KNN Regressor demonstrated the highest adjusted R^2 on both the test set and the final test set. Therefore, the KNN Regressor was chosen as the best model for predicting `overall_rating` due to its superior performance in capturing complex relationships.

8.2 Categorical Outcome: Predicting `player_role`

For the categorical outcome `player_role`, we split the dataset similarly and evaluated several classification algorithms. We couldn't run Random Forest and XGBoost classifiers because they took over 25 minutes to run, even with PCA. We settled for the Decision Tree Classifier since it had a high R^2 compared to those that took so much time to run and had more than enough performance.

For the categorical outcome, we stratified the splits to ensure balanced representation of each player role in both training and test sets. Furthermore, we employed grid search with 10-fold cross-validation using a logistic regression model configured with the *class_weight* parameter set to "**balanced**." This parameter automatically adjusts the weight of each class based on its frequency in the dataset, assigning higher weights to underrepresented classes and lower weights to overrepresented ones. We also set the scoring metric to the weighted F1-score, ensuring that the optimization process prioritized balanced precision and recall across all classes.

Logistic Regression The Logistic Regression model is suitable for both binary and multiclass classification problems. It is interpretable and provides probabilities for class membership, making it useful for understanding the likelihood of different classes. The best model, with $C = 0.01$, achieved a test accuracy of 0.7406 and a final test accuracy of 0.7398, showing consistency in performance. The classification reports indicate strong performance for goalkeepers with near-perfect precision, recall, and F1-scores of 0.99 and 1.00, but somewhat weaker results for midfielders, with an F1-score of 0.43 on the test set. The F1 weighted average for the final test set was 0.76, reflecting an overall good performance, though the model struggled with less frequent classes, especially midfielders.

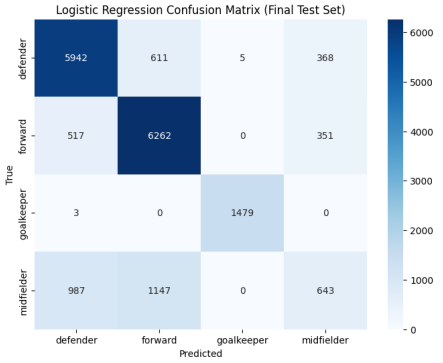

Figure 18: Logistic Regression Confusion Matrix

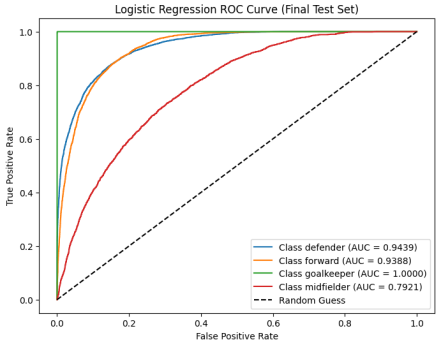

Figure 19: Logistic Regression LROC Curve

Figure 20: Logistic Regression Results

Decision Tree Classifier can handle both numerical and categorical features and capture complex interactions within the data. It is interpretable and performs well with non-linear relationships. The best model, with *max_depth* = None, *min_samples_split* = 2, and *min_samples_leaf* = 1, achieved high accuracy on both the test set and final test set, with a test accuracy of 0.8683 and a final test accuracy of 0.8696. The classification reports show solid performance across player roles, with goalkeepers achieving perfect scores (precision, recall, and F1-score of 1.00). The model showed slightly lower performance for midfielders, with an F1-score of 0.69 on both the test and final test sets. The F1 weighted average scored 0.87 on the final test, indicating a well-balanced performance across all classes.

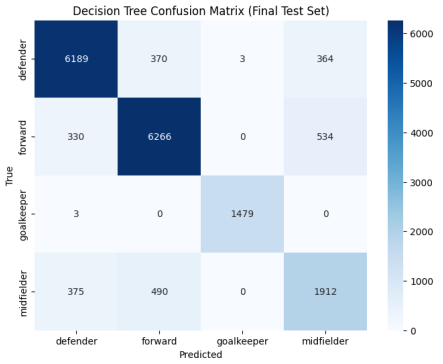

Figure 21: Decision Tree Classifier Confusion Matrix

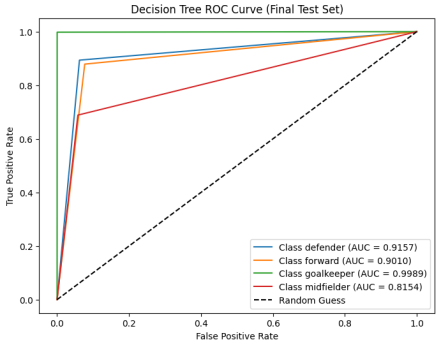

Figure 22: Decision Tree Classifier LROC Curve

Figure 23: Decision Tree Classifier Results

K-Nearest Neighbors (k-NN) Classifier The k-Nearest Neighbors (k-NN) Classifier is suitable for both binary and multiclass classification problems. It is interpretable and provides probabilities for class membership. The best model achieved high accuracy on both the test set and final test set, with a test accuracy of 0.9046 and a final test accuracy of 0.9035. The detailed classification reports show strong performance across all player roles, with the model performing especially well for goalkeepers (perfect precision, recall, and F1-score of 1.00). The F1 weighted average for the final test set was 0.90, demonstrating the model's ability to handle class imbalances effectively. The best hyperparameters were $n\_neighbors = 10$, $p = 2$, and $weights =$ 'distance', which contributed to its high performance across the different player roles.

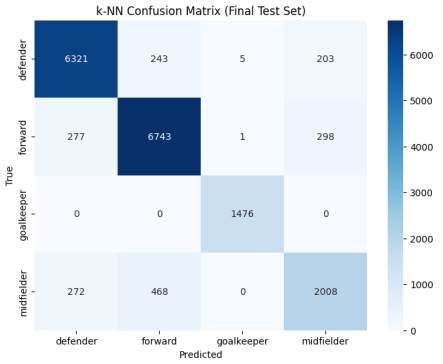

Figure 24: kNN Classifier Confusion Matrix

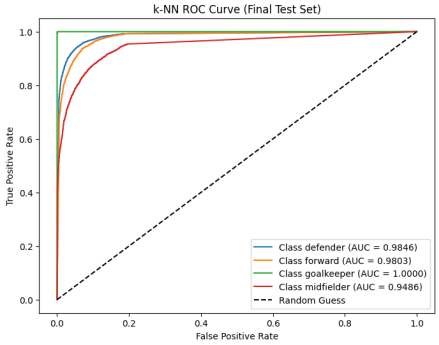

Figure 25: kNN Classifier LROC Curve

Figure 26: kNN Classifier Results

Conclusion for Classification Among the classification models evaluated, the K-Nearest Neighbors (k-NN) Classifier demonstrated the highest F1 weighted average score on both the test set and the final test set. Therefore, the k-NN Classifer was chosen as the best model for predicting `player_role` due to its superior performance in handling complex multiclass classification problems.

8.3 Feature Importance

We conducted feature importance analysis for the chosen regression and classification models. For the classification model K-Nearest Neighbors (k-NN), the components had roughly the same importance, and we did not choose to eliminate any.

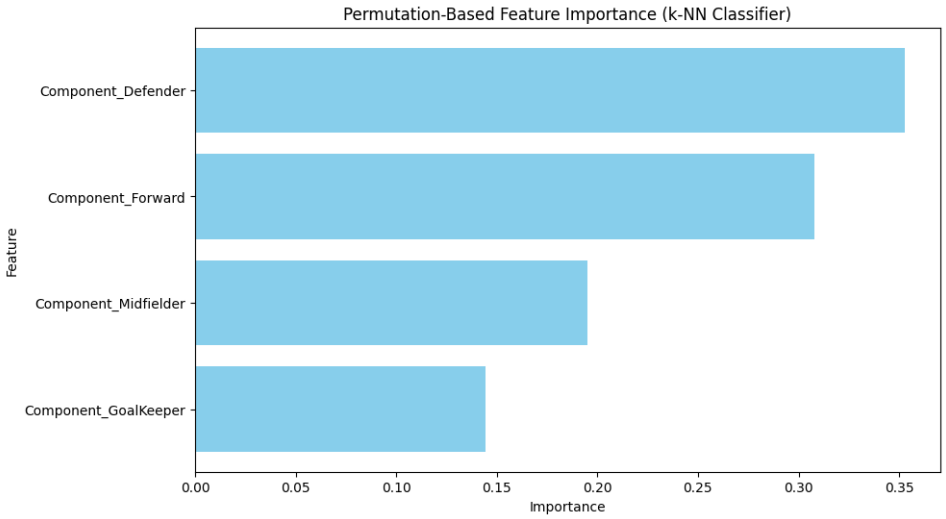

Figure 27: Decision Tree Classifier Feature Importance VIA Permutation

For the regression model (KNN Regressor), we removed *Component_Midfielder* based on its lower importance in permutation feature importance. We then re-trained the k-NN Regressor with the remaining components (*Component_Forward*, *Component_Defender*, and *Component_GoalKeeper*).

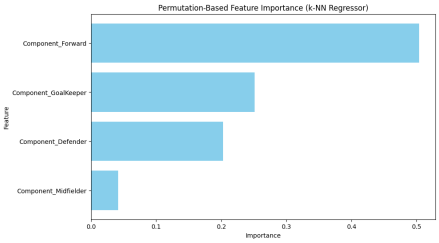

Figure 28: KNN Regressor Feature Importance VIA Permutation

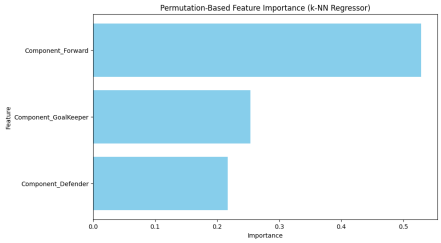

Figure 29: KNN Regressor Feature Importance Post Permutation

Figure 30: KNN Regressor Feature Importance

The optimized k-NN model with $k = 5$ achieved a cross-validation RMSE of 2.2289, a test set adjusted R^2 of 0.9032, and a final test set adjusted R^2 of 0.9043. This feature selection improved model's interpretability and computational ef-

ficiency. The results showed that *Component_Forward* was the most important feature, followed by *Component_GoalKeeper* and *Component_Defender*.

9 Business Implications

Fair Talent Identification: 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 leveraging quantifiable data, the methodology ensures that the best talent is identified based on performance metrics rather than subjective opinions, which can often be influenced by biases. This not only enhances the fairness of the selection process but also increases the likelihood of identifying hidden talents that might otherwise be overlooked.

Optimized Role Assignment: By tailoring role assignments based on key attributes, the methodology enhances team performance. 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, as players are assigned roles that best suit their skill sets.

Scalability for Clubs: The framework's adaptability makes it a valuable tool for talent identification in youth academies and semi-professional clubs worldwide, improving scouting efficiency. The methodology can be scaled to different levels of competition, from amateur to professional, making it a versatile tool for talent scouting. Youth academies and semi-professional clubs can benefit

significantly from this approach, as it provides a structured and objective way to identify and develop young talent, potentially leading to more successful transitions to higher levels of competition.

10 Areas for Further Development

Real-world Validation: The study relies on a proxy dataset of professional football players rather than actual Emlyon students, which limits the direct applicability of the results. While the dataset provides a robust foundation for modeling and analysis, real-world validation is necessary to ensure that the findings are applicable to the target population. Future work should focus on collecting data from Emlyon students and validating the framework in real-world scenarios to enhance its practical utility.

Player-Specific Dynamics: The methodology does not account for intangibles like teamwork, leadership, or adaptability, which are critical in real-world settings. These soft skills are essential for a team's success but are challenging to quantify and incorporate into a data-driven model. Future research could explore ways to integrate these intangible factors into the analysis, providing a more holistic evaluation of players.

Scalability Challenges: The computational intensity of some models (e.g., Random Forest, XGBoost) could hinder scalability for larger or less-resourced organizations. While these models provide valuable insights, their computational requirements may be a barrier for smaller clubs or organizations with limited resources. Addressing this challenge could involve developing more efficient algorithms or leveraging cloud-based computing solutions to make the framework more accessible.

Limited External Variables: The analysis excludes environmental factors like coach strategy or team dynamics, which influence player performance significantly. These external variables play a crucial role in a player's performance and should be considered for a comprehensive evaluation. Future work could focus on incorporating these factors into the model to provide a more accurate prediction of player performance and role suitability.

11 Conclusion

The study presents a robust, data-driven methodology for player selection and role assignment in football, achieving fairness and performance optimization. While the use of advanced machine learning techniques provides valuable insights, real-world validation and consideration of external factors are necessary for broader adoption. Future work could focus on integrating dynamic variables, reducing computational overhead, and validating the framework in real-world scenarios to enhance its practical utility.

Our comprehensive analysis illustrates that a data-driven approach to player selection and role assignment is not only feasible but also advantageous. By leveraging advanced statistical techniques and machine learning models, we can make informed decisions that enhance fairness and optimize team composition. The models developed, particularly the k-NN regressor, demonstrated high predictive accuracy, validating the effectiveness of our methods.

This approach holds significant potential beyond Emlyon Business School. Football clubs globally can adopt and scale this methodology for talent identification among young players. By utilizing data-driven models, clubs can objectively assess player potential and suitability for specific roles, leading to more strategic recruitment and development.

For practical implementation, we recommend utilizing the k-NN regression

model for predicting player performance, given its superior accuracy in capturing non-linear relationships. Addressing class imbalance in classification models is crucial, and we suggest exploring techniques like resampling, adjusting class weights, or using specialized algorithms to improve minority class predictions.

Continuous data collection and model refinement will further enhance predictive capabilities. Incorporating qualitative assessments alongside quantitative models can provide a holistic evaluation of players. Real-world validation through pilot testing and performance monitoring will be essential to fine-tune the models and ensure they generalize well to the student population and beyond.

By adopting this data-driven methodology, Emlyon Business School can assemble a football team that not only embodies excellence on the field but also reflects the institution's commitment to innovation and fairness. Moreover, the scalability of this approach presents an opportunity for football clubs to revolutionize talent scouting and team building in the sport.

12 References

References

FIFA. Fifa dataset, 2024. URL https://www.fifa.com. Accessed: 2024-11-30.

Kaggle. Fifa player dataset on kaggle, 2024. URL https://www.kaggle.com/datasets. Contains player ratings, attributes, and historical performance. Accessed: 2024-11-30.

OpenAI. Chatgpt: A conversational ai model, 2024. URL https://www.openai.com/chatgpt. Accessed: 2024-10-12.

Sports Reference, LLC. Fbref: Advanced football stats and history, 2024. URL https://fbref.com. Accessed: 2024-11-30.

Understat.com. Understat: Expected goals (xg) in football analysis, 2024. URL https://understat.com. Accessed: 2024-11-30.

WhoScored.com. Whoscored: Football statistics and analysis, 2024. URL https://www.whoscored.com. Accessed: 2024-11-30.