The Science of Scoring: Evaluating Efficiency Among Europe's Top Strikers Phase 2

Mohamed Ali Larbi Daho Bachir
Computer Science Major
Undergraduate
University at Buffalo

Aditya Kumar Dwibedi
Computer Science Major
Graduate
University At Buffalo

I. Introduction

Building on our findings and analyses from Phase 1, we aim to present six algorithms that highlight the primary objective of our project: identifying the best-performing forwards in Europe's top five football leagues. For this analysis, we will employ the following algorithms: Linear Regression, K-means clustering, K-Nearest Neighbors (KNN) Classification, and Support Vector Machine (SVM) Classification, along with two additional techniques.

These models will yield critical insights that managers, coaches, and scouts will consider when determining their team's key striker. Two essential factors will be examined through this analysis:

1. Goal-Scoring Proficiency: This attribute is vital for forwards, as it directly correlates with their effectiveness on the field. A striker's ability to convert scoring opportunities is the primary metric by which their performance is often evaluated.

2. Player Position Analysis: This analysis is particularly useful for players who may not be traditional goal scorers but can contribute effectively through playmaking, dribbling, etc. Understanding these dynamics allows coaches to identify potential alternatives for underperforming players, especially if they are deployed out of their optimal position or if they might excel in a different role.

By using these algorithms we develop a handy analysis tool for managers, coaches, and scouts who can now have a much more statistical approach to position-based decision-making.

II. Exploratory Data Analysis

A. 1. Linear regression

As part of our first analysis, we apply a Linear Regression model to analyze the relationship between goals scored and shots on target by forwards. In the diagram below, each point represents a player, with the x-axis showing the number of goals scored and the y-axis representing the number of shots on target. The red line denotes the line of best fit, depicting the linear relationship between the two variables.

As expected, the positive slope of the line indicates that players who take more shots on target tend to score more goals. However, the red line also tells us that players positioned below the line, especially those farther to the right, exhibit a more efficient conversion rate. They score more goals with fewer attempts and therefore these players are particularly valuable to managers and club owners, as they can maximize their scoring ability with fewer attempts.

On the other hand, players positioned above the line, particularly toward the left, represent less efficient forwards. These players take a higher number of shots without an increase in goals produced, making them less ideal in terms of scoring efficiency. Given that, this analysis focuses exclusively on forwards and attacking players, players in this category may deserve a further look into their role and position while keeping the scrutiny of the team's offensive strategy in mind.

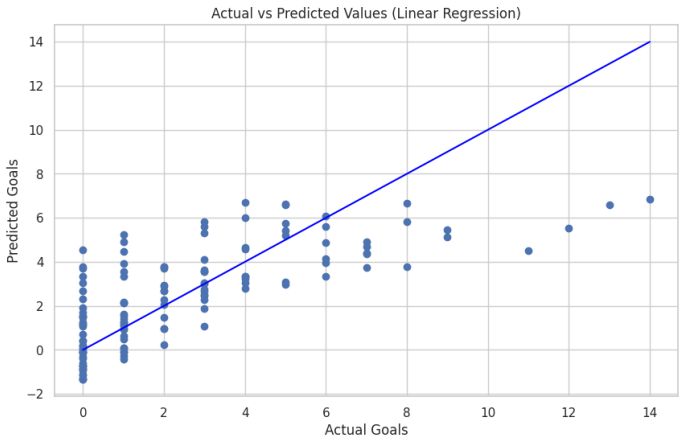

Fig. 1. Linear Regression

B. 2. SVM Analysis

In football, selecting the most efficient players is crucial for optimizing team performance. Both managers and coaches must evaluate numerous statistics—such as goals scored, assists, and minutes played—to determine which players should be on the field. This decision-making process is often complex and tedious. However, machine learning algorithms like SVM analysis provide a powerful solution by using data to predict player performance, particularly in terms of goal-scoring ability.

In our analysis, we apply SVM to predict the number of goals players will score based on their statistical performance graphed against their actual output. The graph below shows

us a player's = predicted goals versus actual goals, with a regression line demonstrating the model's accuracy, achieving an R^2 score of 96 percent. The closeness of the data points to this line indicates a high level of prediction accuracy, reinforcing the model's reliability.

This analysis can be highly valuable for clubs during transfer windows, as it identifies players whose actual goal-scoring ability closely aligns with their potential. By highlighting standout performers, the model helps managers and coaches in making better decisions. While football involves more than just statistics, having a data-driven tool that reveals the true scoring potential of players offers a more precise understanding of their impact on the pitch, helping teams strategically strengthen their lineup.

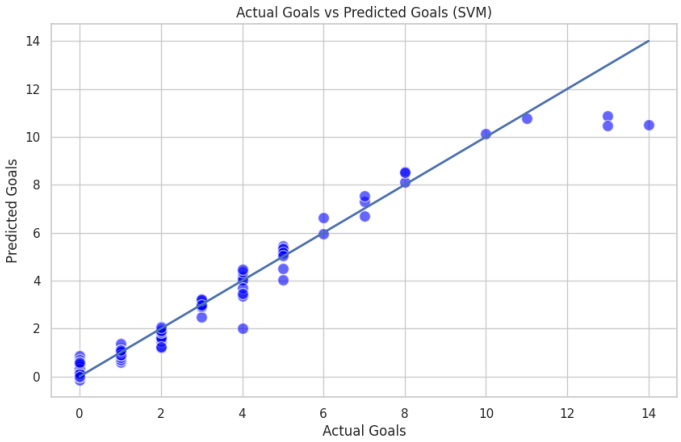

Fig. 2. SVM (Support Vector Machine)

C. 3. XG Boost

As part of our third algorithm we used XG Boost to highlight an offensive player's playmaking ability. Forwards have just as much of a duty to create chances as they have to take them. So we plotted predicted assists vs actual assists. This highlights how likely a player who is predicted to assist a certain number of goals is to actually create those chances irl. Our accuracy score had a R^2 score of 84 percent. This shows a high level of accuracy with a few outliers. The players closest to the line are accurately predicted and those farther away are clearly anomalies.

However upon further analysis we also saw that our model achieved a MSE score of 0.01 which is a clear sign of overfitting. This means that the model may be too tuned to the training data and could struggle with new, unseen data.

For managers and coaches, this model provides valuable insights as for which forwards are consistently reliable in creating chances. By identifying players who are predicted to perform well in assists and actually deliver on the field, managers can make more informed decisions when setting up offensive strategies or scouting talent. Overfitting concerns also highlight the need for refinement to ensure robustness across different match scenarios.

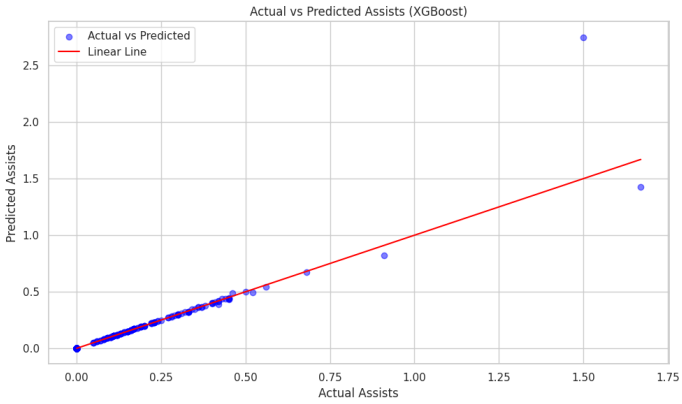

Fig. 3. XGBoost (Extreme Gradient Boosting)

D. 4. K-Means Clustering

In our analysis, we implement the K-Means Clustering algorithm on our dataset, focusing on the relationship between goals scored and minutes played by each player. We initially selected k=3 arbitrarily to explore how the data clusters and the resulting plot form three distinct groups: Group 0, Group 1, and Group 2. Based on the clustering, we can speculate that players in the purple cluster likely represent regular starters, as they tend to play longer and score more goals. Conversely, players in the yellow cluster appear to be substitutes or bench players, as their playing time is significantly lower, and they score fewer goals. This insight could be particularly valuable for team managers, providing data-driven support to determine which players might deserve more playing time based on their potential contributions.

Additionally, we performed an analysis to determine the optimal k value as shown in the second graph. After iterating through multiple values of k, we found that k=3 remains the most suitable choice, as the majority of players still form three clusters, with any additional clusters representing marginal or less significant groups. Increasing the number of clusters beyond three could risk overfitting, as the clustering pattern does not meaningfully improve beyond this point.

E. 5. KNN Classification

In our third analysis, we implement the K-Nearest Neighbors (KNN) classification to categorize players based on their optimal playing positions which we determine by their scoring efficiency and playmaking abilities. We strategically selected k = 5 to align with our data cleaning process, which focused exclusively on five attacking positions, given that our data focuses solely on forwards. By applying KNN, we aim to label each player according to the style or position that best suits their skill set. Logically, players with similar attributes and capabilities will cluster closely together in the feature space. This clustering enables us to make informed predictions about the position or playstyle that a new data point (player) should adopt based on their contributions to the team's offensive dynamics. As illustrated in the graph below, players are

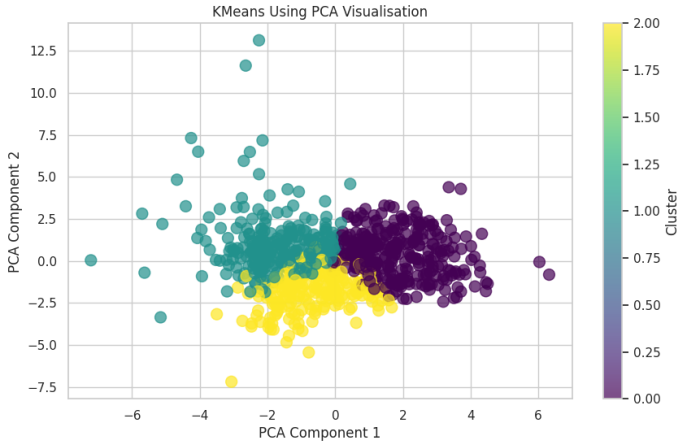

Fig. 4. PCA Visualization (2D)

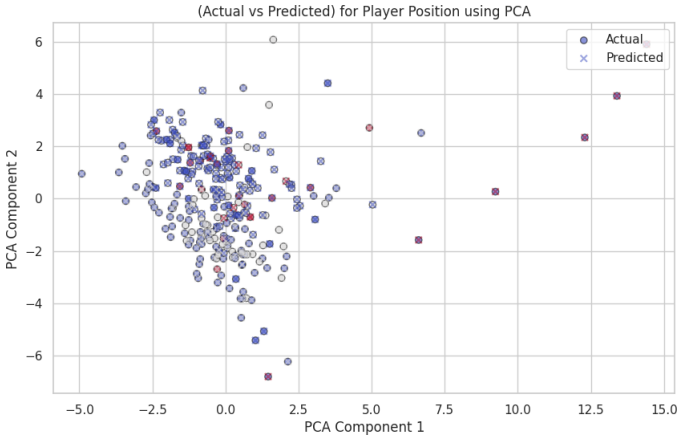

Fig. 6. KNN Classification

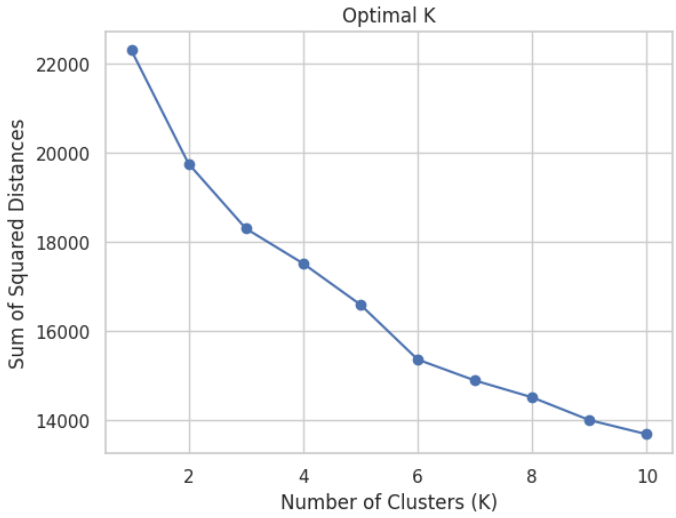

Fig. 5. Picking up the optimal K-value)

positioned according to their attributes, showcasing how those with similar skill sets group together. This visualization not only highlights the distribution of player profiles within each position but also helps analyze how emerging players can be effectively integrated into the team based on their offensive contributions and playstyles.

F. 6. Decision Tree

As part of our own data driven algorithms we chose to apply a decision tree based algorithm which we thought was particularly useful for handling categorical data, like player positions. It can capture interactions between features, such as the relationship between goals scored and defensive contributions. As shown below The Decision Tree model achieved an overall accuracy of 47.85 percent. This is particularly low but it gives us some key information in certain positions. From the graph we can see a huge difference in how the model processes different positions. This can be seen via the large number of overlaps present in positions FW and FWMF. This shows us that the model performed relatively well when predicting forwards, achieving a precision of 0.59 and a recall of 0.62. This means that 59percent of the players predicted as forwards were actually forwards, and the model correctly identified 62 percent of the actual forwards in the dataset. However in the other hand our algorithm failed to predict defensive forwards as there was a precision score of 0.00 which shows that this model was difficult for the model to predict. This is also evident as the data points are widely separated with little to no overlapping. This model helps managers identify key players for attacking roles. However, the model struggled with complex roles on this particular data as it did with Defensive Forwards. This suggests the need for further analysis or refined models, allowing managers to better understand player strengths to make informed tactical decisions.

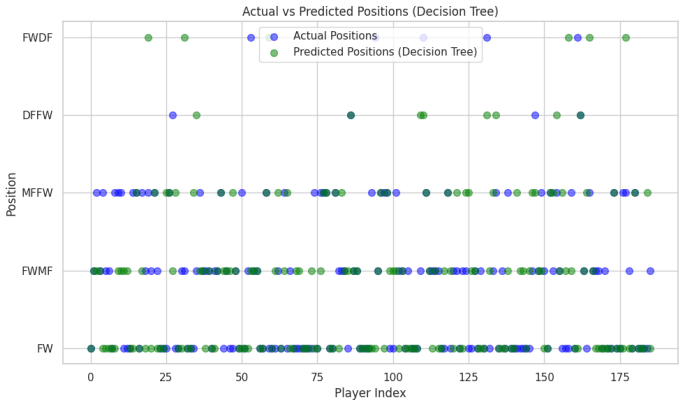

Fig. 7. Decision Tree

G. Index

ACKNOWLEDGMENT

This report was made by an equal effort between Mohamed Ali Larbi Daho Bachir (50366050) and Aditya Kumar Dwibedi (50347861).

REFERENCES

[1] https://www.kaggle.com/datasets/vivovinco/20222023-football-player-stats
[2] https://www.coursera.org/articles/machine-learning-algorithms
[3] https://xgboost.readthedocs.io/en/latest/tutorials/model.html
[4] https://scikit-learn.org/1.5/modules/tree.html
[5] https://scikit-learn.org/dev/modules/generated/sklearn.decomposition.PCA.html