Predicting Football Player Market Values in Europe's Top 5 Leagues (2019-2023)

Timothy Cassel, Tirdod Behbehani

December 20, 2024

Contents

1 Introduction

In this research, we aim to predict the market values of football players in the top five European leagues from the 2019/20 through the 2022/23 seasons. Our investigation focuses on two key objectives. First, we seek to determine how a range of factors influence player valuations and to what degree these characteristics shape their value on the open market. Second, we aim to identify and group player attributes into distinct clusters, potentially identifying particular player profiles that may shed some insight into their valuation.

In Section 2, we give a brief overview of related research and provide background on a couple of similar studies. In Section 3, we present a detailed overview of our dataset. Specifically, we use valuation data from Transfermarkt, combined with players' performance metrics from FBref, resulting in two final datasets—one featuring per-90-minute statistics and another containing total statistics. We ultimately focus on the totals dataset. Our dependent variable is the Transfermarkt crowd-sourced market value, while our input consists of 30 selected player attributes. Our focus is on including predominantly player performance statistics rather than other market value determinants.

In Section 4, we introduce the penalized/regression methods employed: OLS (baseline), LASSO, Ridge, Elastic Net, and Adaptive LASSO—and discuss their results. Our findings indicate that all methods perform similarly, with Adaptive LASSO having a marginal advantage. This suggests that penalized likelihood may not improve predictive performance over OLS for our dataset. Given a larger dataset with more variables, it is possible that the performance would improve, and the penalized regression methods would properly set non-important variables to zero, and identify the best subset among higher dimensionality.

In Section 5, we explore unsupervised learning techniques such as K-Means Clustering, PCA, and Sparse PCA. We found that K-Means Clustering did not do a good job of differentiating between clusters, with heavy overlap amongst different clusters. Sparse PCA performed marginally better than PCA, requiring one less critical component to explain 75% of the cumulative variance. Furthermore, Sparse PCA was easier to interpret due to its inherently focused components. Sparse PCA was able to group players into buckets based on creativity, defensive prowess, goal scoring ability, and age. Interactions between those principal components can provide deeper insights into players' style and ability.

2 Related Work

Franceschi (2023) [4], offers an in-depth literature review on the determinants of football player valuations. Market value prediction can be done using estimated market values from Transfermarkt estimates (or other sources), or by the actual transfer fees paid. Transfermarkt offers subjective, crowd-sourced player market values that reflect an "expected free-market value," which are distinct from the actual transfer fees, but have been shown to be empirically related [4], though often biased [3]. Among others, [5] uses LASSO regression to predict market values of LaLiga players and strikers. This work aims extend this work to include more positions and leagues.

While we could find only limited research on clustering of football players, [6] use a combined methodology of k-means clustering and particle swarm optimization to identify 3 distinct clusters of players; players who can play 3 positions, players specialized in one position, and players who can play 2 positions. We do not use this more advanced methodology but aim to see if we can identify different clusters mainly based on player performance statistics.

3 Dataset

3.1 Overview

We use two Kaggle datasets: one containing player market values scraped from Transfermarkt [2], and another with player performance statistics scraped from FBref [1].

We had to do a substantial amount of preprocessing to integrate the data from our two data sources. We created two versions of the player performance data: a "Totals" dataset containing raw cumulative season statistics, and a "Per 90s" dataset, which normalizes these statistics by "90-minutes" played. While we initially had over 90 very specific player statistics variables, we ultimately reduced our feature set to 30 to avoid excessive multicollinearity.

Merging player valuations and performance data required joining on league, club, season, and player name. Due to inconsistencies in player names, we applied fuzzy matching with a 75% similarity threshold, resulting in 144 players who could not be matched between sources.

We excluded goalkeepers and players with fewer than 500 playing minutes to generate a more representative, "active player" sample. Missing values mainly were due to statistics not accumulated by a player (e.g., no goals scored all season); these were replaced with 0. The final dataset contained 7,181 observations from 3,049 players across 129 teams. After a 70/30 train-test split, we had 5,026 observations in the training set and 2,155 in the test set.

3.2 Dependent and Independent Variables

Our dependent variable is the log of the season-end market value in euros. Using the log transformation reduces the right-skewness of the data and approximates a normal distribution, although it complicates direct interpretation of the coefficients, since they now reflect percentage changes rather than absolute changes. The summary statistics for this are shown below in Table 1. From this, it is also evident that market values such as 180 million will be difficult to predict with a linear model.

Table 1: Summary Statistics for Market Values (Original and Log-Transformed)

Dataset	Scale	Mean	Median	SD	Min	Max
Training	Original	12,839,649	6,500,000	17,030,593	25,000	180,000,000
Test	Original	12,680,452	6,500,000	16,256,856	25,000	150,000,000
Training	Log	15.67	15.69	1.25	10.13	19.01
Test	Log	15.68	15.69	1.22	10.13	18.83

Initially, we aimed to include comprehensive player performance statistics, but interpretability suffered due to multicollinearity. We therefore selected 30 key variables for both the Totals and Per 90s datasets. Among these are control variables such as age, age^2 (reflecting the non-linear age-value relationship [4]), one-hot encoded season and league indicators, and a binary variable indicating whether a player is from a top-10 footballing nation (according to current FIFA rankings). Team effects were captured using target encoding (dimension=128) to avoid introducing a large number of dummy variables.

Key performance metrics include Goals, Assists, Shots, Playing Time, Expected Goals, Expected Assists, Progressive Ball Carries, Progressive Passes, Key Passes, Tackles, Interceptions, and Errors. While some of these measures are correlated, we rely on penalized regression and dimensionality reduction methods to help isolate the most relevant factors.

We acknowledge that omitting certain club or contract-related variables (such as wage and contract length), known to influence market value, is a limitation. Other potential correlates of market value are also not considered.

4 Regression

We approach this regression problem using the following methodology. We start with a simple baseline OLS model, and run 4 other models (Lasso, Ridge, Elastic Net, and Adaptive Lasso). To assess model performance, we consider the R^2, and $RMSE$, as well as the BIC. We choose to consider BIC because we care more about interpretability of the model, rather than prediction. We do not want an overly complex model, if we cannot interpret the parameters.

To establish a baseline result, we first ran a simple linear regression using ordinary least squares, which we subsequently use to compare to each of our penalized regression methods. We use cross-validation for variable selection, since we have already have a relatively low-dimensional dataset, relative to its size.

4.1 Methodologies

4.1.1 Lasso

Lasso is a type of linear regression that adds an l_1 penalty to the least squares objective function. The objective function for Lasso can be expressed as follows:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(\log(y_i) - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda \|\boldsymbol{\beta}\|_1,$$

4.1.2 Adaptive Lasso

Adaptive Lasso is an extension of Lasso regression, which uses a modified version of the penalty term. In Lasso, all coefficients are penalized equally, however, Adaptive Lasso introduces a weighted penalty. Features with larger initial estimates (indicating greater importance) receive smaller penalties due to their smaller weights, while features with smaller initial estimates are penalized more heavily. In our version of lasso, we use the estimates from the previous Lasso model as the weights. The objective function for Adaptive Lasso can be shown as below:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(\log(y_i) - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{\hat{w}_j},$$

4.1.3 Ridge

Ridge regression is an alternative type of penalized regression method that adds an l_2 penalty to the LS objective function. Ridge regression penalizes all coefficients the same, shrinking them toward zero. However, it does not set any coefficients exactly to zero, meaning Ridge regression retains all predictors in the model. This makes it more effective in the presence of multicollinearity, which is why we expect ridge regression to possibly perform well in our model, as we have correlated features.

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(\log(y_i) - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda \|\boldsymbol{\beta}\|_2^2,$$

4.1.4 Elastic Net

Elastic Net is a combination of both ridge and lasso regression, adding both penalty terms, and is most ideal in setting with high-dimensional data. Though this is not the case for our model, we try using this as well.

4.2 Results

Table 2: Performance of various models on the Totals Dataset (Market Value is log-transformed)

Model	**RMSE (log)**	**R² (log)**	**BIC (log)**
OLS (log)	0.6534	0.7108	10571.790
Ridge (log)	0.6685	0.6972	-3434.162
Lasso (log)	0.6534	0.7108	-3698.828
Elastic Net (log)	0.6534	0.7107	-3698.897
Adaptive Lasso (log)	**0.6528**	**0.7113**	**-3755.039**

Table 3: Performance of various models on the Per 90s Dataset (Market Value is log-transformed)

Model	**RMSE (log)**	**R² (log)**	**BIC (log)**
OLS (log)	0.6462	0.7171	10473.238
Ridge (log)	0.6602	0.7047	-3532.034
Lasso (log)	0.6462	0.7171	-3797.360
Elastic Net (log)	0.6462	0.7171	-3797.441
Adaptive Lasso (log)	**0.6460**	**0.7173**	**-3854.222**

In Table 3 we can see the RMSE (log), R^2 (log), and BIC (log) of each of the 5 models tested on the *Totals* dataset. We can see that the performance of each of the models is extremely similar on RMSE, and R^2 (log). The RMSE (log) suggests that the adaptive lasso regression model is marginally better than the others. The BIC is lowest for the Adaptive Lasso and hence also favors this model. Finally, an R^2 value of approximately 71.1% suggests that the covariates that we include are capable of explaining 71,1% of the variation in player market value.

This is logical, since we already know that there are determinants of the player value, which we have not accounted for, and so do not expect to get a much higher value. Since in our models, the penalization methods, do not set any variable coefficients to zero, it could be interpreted that in this case OLS might be favourable as long as its assumptions hold, because inference on OLS is simpler.

However, for this dataset, this may not be the case, especially due to the presence of heteroskedastic errors. In further iterations, we could think about adding more variables to the model to add dimensionality, which may allow the penalization methods to be more selective. Similarly, Table 4 shows the same results for the 90s dataset. Since results are very similar, we will just use one of the datasets, i.e. the totals one for simplicity of interpretations.

Next, we consider the insights we can gather from the coefficients of our best model, the Adaptive Lasso.

The coefficients in Table 4 should be seen as associations, rather than causal effects. Doing inference on these results requires a more involved approach, which we do not address here. However, we can consider the effect sizes, and see if they align with our expectations. Age showed a strong positive relationship with market value (coefficient: 1.37), and the negative coefficient for its squared term (-1.88) indicates diminishing returns as players age. These effects are likely so large since they partially counteract each other. These results are expected and in line with past research, indicating that age is a positive indicator of market value, however, this effect diminishes and reverses as players fall from their peak.

Key performance metrics—such as playing time (0.14), goals scored (0.11), and expected goals(xG) (0.10) exhibited positive associations with market value (e.g. one additional goal scored, is associated with a 14% increase in their market value). Similarly, progressive passes (0.12), assists (0.04), have positive associations. Additionally, players from good teams seem to benefit a lot, in regards to their market value. Even though it is difficult to interpret the exact effect size of the encoded team variable, we can say that a one-unit increase in this variable is correlated with a 65% increase in a player's market value. It would be interesting to compare these results to the coefficients of the OLS regression, to get a better understanding of which variables are statistically significant in that model, and if the effect sizes are also the same. Perhaps surprisingly, defensive stats have almost no bearing on market value. This might suggest that defensive players should also be separated into separate datasets or models, to account for differences in player attributes and statistics. Additionally, for season variables, we do see that later seasons have a slightly higher effect, which aligns with expectations of inflation of market values over time.

Table 4: Adaptive Lasso Selected Predictors of Football Player Market Value (Log Scale)

Predictor	**Coefficient**
Age	1.37
Age Squared	-1.88
Playing Time (Minutes)	0.14
Goals Scored	0.11
Expected Goals (xG)	0.10
Performance (Fouls Drawn)	0.00
Aerial Duels Won	0.05
Successful Take-Ons	0.09
Tackles Won	-0.04
Clearances	0.10
Assists	0.04
Progressive Passes	0.12
Top 10 Nation	-0.01
Forward Position (FW)	0.11
Midfield Position (MF)	0.11
Season 2020/2021	0.02
Season 2021/2022	0.03
Season 2022/2023	0.06
League: English Premier League	0.05
League: La Liga	0.02
League: Ligue 1	-0.04
League: Serie A	0.03
Team Encoded	0.65

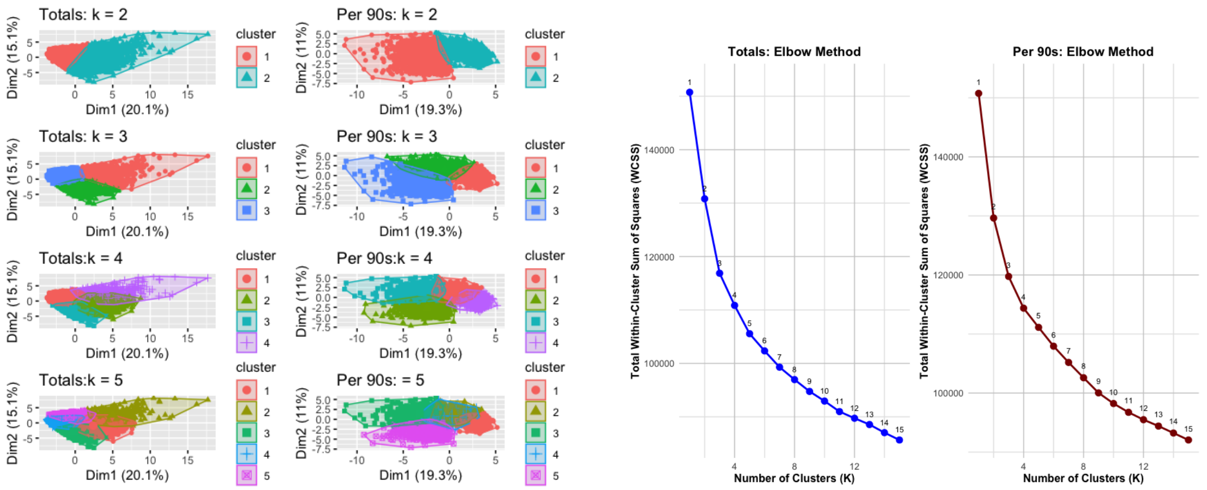
5 Unsupervised Learning

5.1 Overview:

Our goal in this section is to assess whether unsupervised learning methods can effectively group our input features into well-defined clusters. In this section, we will assess the performance of K-Means Clustering, Principal Component Analysis, and Sparse Principal component analysis. Our K-Means Clustering found that adding clusters to our dataset did not significantly improve the clustering fit. This led us to explore reducing dimensionality through PCA and Sparse PCA.

5.2 K-Means Clustering

K-Means Clustering is an unsupervised learning algorithm which seeks to group data into a predetermined number of clusters, doing so using feature similarity. The hope was that it would uncover meaningful insights, perhaps showing that player position, goal scorers, or dynamic dribblers could be put into clusters. We performed K-Means Clustering on the totals and per 90s training datasets. We set K = 2, 3, 4, and 5.



(a) K-Means Clustering on the Totals and Per 90s Datasets for $k = 2, 3, 4,$ and 5.

(b) Elbow Method on the Totals and Per 90s Datasets.

Figure 1: Comparison of K-Means Clustering and the Elbow Method.

The clusters are not informative. The plots in Figure 1a show significant overlap across clusters, indicating a lack of distinct cluster structures.

We can use the elbow method to try to identify the optimal number of clusters. We examine the within-cluster sum of squares for each cluster centroids to identify the point where adding additional clusters provides a minimal sum of squares reduction.

As we can see in Figure 1b, there is no point in which the within-cluster sum of squares plateaus. This indicates that we have homogeneous data that lacks the variability to be split into distinct clusters. Furthermore, the lack of a distinct elbow may be attributed to the "curse of dimensionality", which reduces the effect of clustering. From this point forward, we focus on the totals dataset, as it is largely similar to the per-90s dataset.

5.3 Principal Component Analysis (PCA)

We can use PCA to address the challenges that we faced while clustering. PCA reduces dimensionality while identifying the principal components that capture the most variance. The objective function for PCA can be expressed as follows:

$$\underset{u_1:\|u_1\|_2=1}{\text{argmax}}\ u_1^\top \Sigma u_1$$

Figure 2 shows us that 74% of the variance is explained with the first 9 principal components, 80% is explained with the first 11 principal components, and 91% is explained with the first 16 principal components. We have 30 features, so we can explain 70-80% of the variance with about a third of the features, but it takes about half of the features to capture 90% of the variance.

PCA has dense loading, which means that many variables have non-zero coefficients for each principal component. Interpretation of PCA can be challenging due to an overlap of similar variables. For instance, PC1 features key passes, progressive passes, assists. To address this, we can reduce dimensionality and improve interpretability.

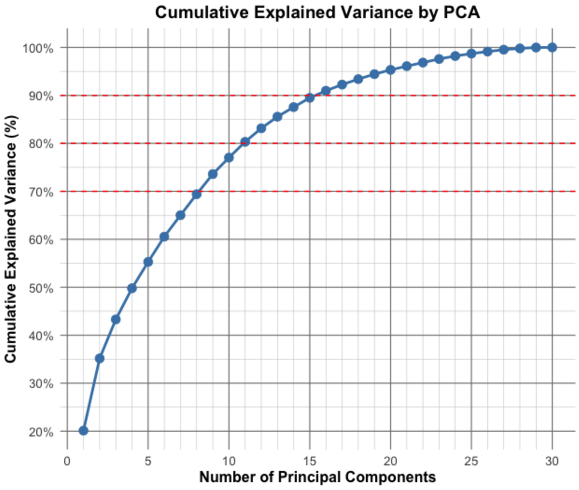


Figure 2: PCA on the Totals Dataset

5.4 Sparse Principal Component Analysis

While reducing dimensionality, we do not merely want to explain all of the variance or minimize the error. Instead, we want to strike a balance between identifying different player archetypes, while still being able to explain a reasonable amount of variance. Unlike Standard PCA, Sparse PCA employs an L1 penalty, which encourages some coefficients to shrink to exactly 0. The sparsity simplifies the principal components by focusing on solely the most important variables, which makes them more interpretable.
The objective function for Sparse PCA is as follows:

$$\max_{u_1:\|u_1\|_2=1} \; u_1^\top \Sigma u_1 \quad \text{subject to } \|u_1\|_1 \leq \lambda$$

We used cross-validation to tune the K and λ parameters. A larger K means more principal components, and a larger λ encourages more sparsity. Our cross-validation found that the setting $K = 10$ and $\lambda = 0.05$ explained 75.66% of the cumulative variance. Although continuing to increase K would explain more variance, this would have come at the expense of the ability to interpret the principal components. Figure 4 demonstrates this below.

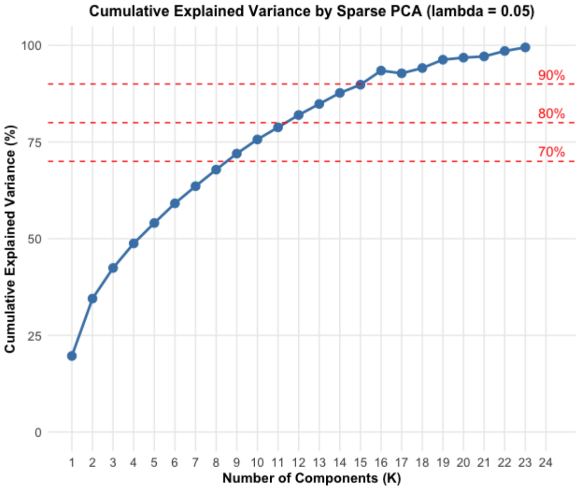


Figure 3: Sparse PCA on the Totals Dataset: $\lambda = 0.05$

We were able to identify some interesting patterns in the first four components with $K = 10$ and $\lambda = 0.05$. Principal Component 1 identifies ”playmakers”, Principal Component 2 identifies ”defenders”, Principal Component 3 identifies ”forwards/goalscorers”, and Principal Component 4 identifies ”player age”. The remaining six clusters, however, are not as interpretable as the first four clusters.

Variable	**PC1**
kp	0.60039467
carries_prgc	0.50731604
take_ons_succ	0.39984109
ast	0.33713962
performance_fld	0.22872520
progressive_passes	0.18322762
xA	0.14798879
aerial_duels_won	-0.02885131

Variable	**PC2**
clr	-0.5349560
int	-0.4729286
blocks	-0.3887240
tackles_tklw	-0.3440809
playing_time_min	-0.3412481
aerial_duels_won	-0.2474920
err	-0.1588005
progressive_passes	-0.1332010

Variable	**PC3**
expected_xg	0.75474896
standard_gls	0.48488396
standard_pk	0.40431323
is_FW	0.13672522
aerial_duels_won	0.10998508
performance_fld	0.03133465

Variable	**PC4**
age	-0.7256311
age_squared	-0.6880840

Table 5: Top variables for the first four principal components when $K = 10$ and $\lambda = 0.05$.

We can combine the principal components and use the loading scores to derive player archetypes. For instance:

- High PC1, High PC2, High PC3, High PC4: Creative, young forwards
- High PC1, Low PC2, Low PC3, High PC4: Creative, young midfielders
- Low PC1, Low PC2, Low PC3, High PC4: Defensive veterans

5.5 Results

While our K-Means Clustering did not perform well, PCA and Sparse PCA show promise in their ability to reduce dimensionality and to identify player archetypes. PCA and Sparse PCA explain relatively similar variance levels, but Sparse PCA's principal components are easier to interpret than PCA. With Sparse PCA, we determined the optimal number of principal components (K) to be 10 and the L1 penalty (λ) to be 0.05. This combination explained 75.66% of the variance in the training data. The interactions between the sparse principal components serve as a starting point for analyzing a player's role, style, and age.

6 Discussion

In this paper, we aimed to predict football player market values within Europe's top five leagues. We wanted to first, assess the predictive power of primary penalized regression methods, and second explore if we could find structures within our data to justify certain clusters of player types, which subsequently could be used to improve results on regression models. In terms of market value prediction, the performance of all regression methods we tested OLS, Ridge, LASSO, Elastic Net, and Adaptive LASSO, was extremely similar.

Even though we expected that penalization methods would help reduce the set of covariates, primarily due to multicollinearity, our dataset consisting of a set of 30 features may have been cut down too much prior to the regressions such that penalized methods were not very different from OLS. Adaptive LASSO showed a marginal improvement in predictive accuracy (RMSE) and had a marginally larger R^2. Leaving out key explanatory factors such as player height, preferred foot, contract durations, player agents, or team-specific financial conditions that can heavily influence real-world transfer fees and perceived market values, may have been a large reason we do not see more accurate predictions, since the crowd-sourced values, our dependent variable, heavily rely on these types of features too.

The regression coefficients of our final model intuitively aligned with prior understanding: younger, but not too young, players with strong offensive metrics and playing time at top-tier clubs were generally associated with higher valuations. Defensive statistics were a relatively minor role, possibly reflecting that our models were biased towards attacking-type players, and overvalued features such as goals, and assists, relative to defensive statistics.

We attempted to use K-Means clustering, however, this did not result in interpretable clusters, Standard PCA showed that a small number of principal components could already explain a large portion of the variance; however, again we found the results to be hard to interpret. Sparse PCA, identified a smaller subset of interpretable principal components, highlighting clear player archetypes such as "creative playmakers," "defensive veterans" and "goal-scoring forwards." An area for future exploration could involve building separate models to predict market valuations for defenders, midfielders, and forwards, leveraging insights from our Sparse PCA guide parameter tuning.

It is very difficult to predict a player's market value. There is a large degree of variance that we were unable to account for, such as player wage, player injury history, and player fame/notoriety. Additionally, TransferMarkt's market value metric is highly volatile in nature, as it is crowd-sourced and reactive to real-life player transfers. Ultimately, the best approach to predicting a player's market value may involve first estimating the player's on-field value, then leveraging latent variable analysis to uncover hidden features which influence a player's market value.

References

[1] beridzeg45. *Player Stats From Top European Football Leagues*. Accessed: 2024-12-01. 2024. URL: https://www.kaggle.com/datasets/beridzeg45/top-league-footballer-stats-2000-2023-seasons/data.

[2] David Cariboo. *Football Data from Transfermarkt*. Accessed: 2024-11-13. 2024. URL: https://www.kaggle.com/datasets/davidcariboo/player-scores.

[3] Dennis Coates and Petr Parshakov. "The wisdom of crowds and transfer market values". In: *European Journal of Operational Research* 301.2 (2022). Discusses crowd-sourced player valuations such as Transfermarkt and their biases as proxies for actual transfer fees., pp. 523–534. DOI: 10.1016/j.ejor.2021.10.046. URL: https://doi.org/10.1016/j.ejor.2021.10.046.

[4] Maxence Franceschi et al. "Determinants of football players' valuation: A systematic review". In: *Journal of Economic Surveys* 38.3 (2023). Available under Creative Commons Attribution-NonCommercial License, pp. 577–600. DOI: 10.1111/joes.12552. URL: https://onlinelibrary.wiley.com/doi/10.1111/joes.12552.

[5] Miao He, Ricardo Cachucho, and Arno Knobbe. "Football Player's Performance and Market Value". In: *Proceedings of MLSA15*. LIACS, Leiden University, The Netherlands; Amsterdam University of Applied Sciences, The Netherlands. 2015.

[6] Adam Fahmi Khariri et al. "Implementation of K-Means Particle Swarm Optimization for Clustering Football Players in the Top Five European Football Leagues". In: *Applied and Computational Mathematics*. Ed. by Dieky Adzkiya and Kistosil Fahim. Singapore: Springer Nature Singapore, 2024, pp. 63–75. ISBN: 978-981-97-2136-8.