Prediction of Football Players' Value in the Transfer Market of Well-known European Leagues based on FIFA 19 and Real-world Data

Yu Sun
International Football Education School
Jilin Agricultural University, China
yusun0111@163.com

Kepeng Gu
International Football Education School
Jilin Agricultural University, China
Kepenggu123098@gmail.com

Abstract: *The study delves into FIFA's role as the global regulatory authority for football, managing the sport's development and major events like the FIFA World Cup. FIFA's influence extends to economic goals, impacting football clubs globally as they invest in skilled players. The market valuation of players is crucial, guiding budget allocation for transfers. Using data from the FIFA 19 video game and real-world statistics, the study employs Decision Tree Regression (DTR) and Random Forest Regression (RFR) models, addressing multicollinearity with Variance Inflation Factor (VIF). The Rhizostoma Optimization Algorithm (ROA) and Dwarf Mongoose Optimizer (DMO) optimize models. Results show RFR-based models, particularly RFRO, outperform DTR-based ones, achieving over 99% R2 value and 12% error relative to mean market values. Ensemble models RFRD and DTRD provide a reliable prediction capability of around 98%, aiding real-world decision-making in the football transfer market for club managers, coaches, and stakeholders across different leagues*.

Keywords: *Transfer market, European football leagues, data engineering, variance inflation factor, regression prediction algorithms, rhizostoma optimization algorithm*.

Received February 21, 2024; accepted July 8, 2024
https://doi.org/10.34028/iajit/21/4/13

1. Introduction

Football is one of the most popular sports in the world [43]. A football league is a structured form of competition in which various football teams compete against each other in home-and-away matches, taking place at both the national and international levels [6]. Undoubtedly, European football stands as the most widely embraced sport globally [33]. The preeminent football leagues in Europe comprise the English Premier League (EPL), Italian Serie A, German Bundesliga, Spanish La Liga, and French Ligue 1, collectively recognized as the big five [48].

The EPL, established in 1992, holds a position of great attractiveness within the soccer community, boasting an extensive international fanbase estimated at around 1.46 billion [46]. The league, with 20 clubs, draws appeal from its competitiveness, historic rivalries, and the attraction of top football talent with high player market values [14].

The La Liga, officially Primera División de La Liga de Fútbol professional, is Spain's premier professional football league, renowned globally for its high level of competition. Teams qualify based on performance or promotion from La Liga smart bank. The league is distinguished by its technical quality and passionate fan base, with a special focus on historic rivals Real Madrid and FC Barcelona, known for their intense El-Clásicorivalry [19, 30].

The Bundesliga, founded in 1962, expanded from 16 to 18 teams per season. Its relegation system changed over time, and since 2008/09, two teams are relegated, while the 16th-place team faces the third-place team from the second Bundesliga in a two-leg match. Ranked 3rd in the UEFA coefficient ranking, the top three Bundesliga teams qualify for the UEFA Champions League, the 4th -place team enters the play-off round. The DFB-Pokal winner secures a UEFA Europa League spot, with the fifth and sixth places in contention for Europa League Play-Offs [21, 47].

The Serie A, the Premier League in Italy, is known for its tactical and defensive style and features iconic clubs like Juventus and AC Milan [7, 41]. With 20 teams, the league follows a home-and-away format during the regular season. Lega Calcio Serie A independently oversees the league, operating under the guidance of the Federazione Italiana Giuoco Calcio (FIGC), the Italian football association, which sets the operational guidelines for the championship [12].

The Ligue 1, the foremost league in France established in 1898 and is characterized by exciting attacking football, with Paris Saint-Germain (PSG) being a dominant force [17, 25].

The market value of a player is an estimate of the money a club would be inclined to invest in acquiring the player, regardless of any real transaction occurring [18]. The assessment of players' financial worth, a key factor in transfer negotiations, has traditionally been the domain of clubs and journalists. However, the rising influence of new technologies and the Internet has elevated the significance of football fans in this aspect [8, 38]. The valuation of football players in the market is a dynamic and intricate concept influenced by various factors such as skill proficiency, age, performance during matches, contractual status, and the demand from potential buyers [32]. Negotiations for transfer fees hinge upon these elements, particularly evident in the substantial sums commanded by elite players [35]. Additionally, external dynamics, encompassing industry trends, economic conditions, and the financial capabilities of the leagues' transfer market, play a pivotal role in shaping player valuations [16].

Some authors utilized crowd-based valuation of players. Crowd-sourced player valuations, commonly sourced from the transfermarkt.com website (TM), involve members providing their assessments, and a panel of experts calculates a weighted average to determine a single transfer value for each player. Although TM values exhibit a strong correlation with transfer fees, as noted by Herm *et al*. [23], the two quantities differ fundamentally (TM values represent subjective crowd assessments, while transfer fees are actual payments between clubs). Coates and Parshakov [13] identified systematic bias in TM valuations, which improved with the inclusion of additional covariates in a regression model on transfer fees.

On the other hand, Machine Learning (ML) algorithms, however, offer enhanced predictive power compared to traditional methods [9]. There is some researched worked on the application of predictive ML algorithms in the case of each of the introduced European leagues. For instance, Wang *et al*. [49] examine diverse predictive modeling techniques such as regression analysis, Neural Networks (NN), and XGBoost in the realm of transfer market dynamics in the EPL. The study focuses on identifying optimal strategies for determining player market value and transfer fees, ultimately highlighting NN as the most effective method for clubs in assessing transfer fees. Putra *et al*. [39] investigate the market value of loan players in the English Premier League. Their objective was to identify the factors influencing a player's market value at the conclusion of the loan period. Sengupta [44] studied and comprehended the correlation between the performance of soccer players and their corresponding values in the European transfer market during the most recent season of La Liga, Spain's premier division.

Additionally, the study addresses the observation that the market value of foreign players tends to be higher when compared to that of local players in the league. Moreover, Lepschy *et al*. [28] analyzed the success of three seasons in the German Bundesliga, utilizing various factors such as defensive errors, market value, goal efficiency, shots on goal, and total shots to assess the outcomes of league games. Horn [24] has explored the key indicators for determining the performance and transfer costs of players in the second division of the German Bundesliga.

According to the reviewed literature, most of the studies worked on a single league dataset and there are a limited number of comprehensive researches through different leagues. For example, Podzemsky [37] studied the correlation between a player's market value and its impact on team success in the Premier League, La Liga, and Serie A using correlation analysis and ordinary least squares regression and Felipe *et al*. [16] investigated factors influencing football player market values in top European leagues, exploring both current and maximum economic values for players with professional contracts in Spain, England, Italy, France, and Germany.

In the past two decades, machine learning has been instrumental in converting football statistics into valuable information, enabling real-time analysis of opponents and enhancing decision-making for teams and coaches. FIFA 19 integrates real-world statistical datasets as one of the most reliable datasets and offers a lifelike representation of football. Developed by electronic arts, the game utilizes official player statistics, team dynamics, and historical match data to enhance realism. This research employs a novel machine learning approach to predict the market value of football players of five different European leagues, presenting an advanced and data-driven methodology for understanding and forecasting player valuation in the football industry. The primary innovation in this article lies in the application of the Variance Inflation Factor (VIF) for selecting the most relevant features and addressing multicollinearity. Also, hybrid forms of two tree-based models (Decision Tree Regression (DTR) and Random Forest Regression (RFR)) optimized with the recently developed Rhizostoma Optimization Algorithm (ROA) and Dwarf Mongoose Optimizer (DMO) introduced one of the most prediction tools and ensembling effectiveness of two optimizers by weighting average approach added reliability and powerfulness to the predictions.

The subsequent sections of the study are organized in a way that in section 2, a description of regression base models and optimization algorithms and their representative pseudocodes and flowcharts have been presented. Section 3 outlines the refinement of the dataset sourced from FIFA 19 and real-world statistics to predict football players' market value. Steps include data engineering to dataset cleaning, ten-fold cross-validation for model reliability, feature selection to address multicollinearity and selecting the most imperative features in predicting market value of football players. Then, in section 4, the prediction results are evaluated utilizing various metrics such as R2, RMSE, MSE, MARE, and NSE to assess the accuracy of developed models in predicting football players' market values. In section 5, the comparative performance of predictors is analyzed and discussed using scatter, error, and Taylor plots. Also, in this section the future directions presented to more enhanced predictions in the football transfer market. Finally, in section 6, all obtained results concluded to give comprehensive insights into the contributions of the study.

2. Regression Models and Optimization Algorithms

2.1. Decision Tree Regression (DTR)

The decision tree serves as a supervised learning approach for regression and classification challenges [27], featuring hierarchical levels or divisions in its structure. In instances where a specific category or class is absent, the regression method can be employed to predict outcomes based on independent variables [2, 15].

A basic decision tree model has a lone binary target variable (Y) and two continuous variables (x_1 and x_2) spanning from 0 to 1. Because DTR analysis seeks to best divide all available data into discrete portions, each segment, or leaf node, in the analysis is directly related to the final output of sequential decision-making processes as shown in Figure 1.

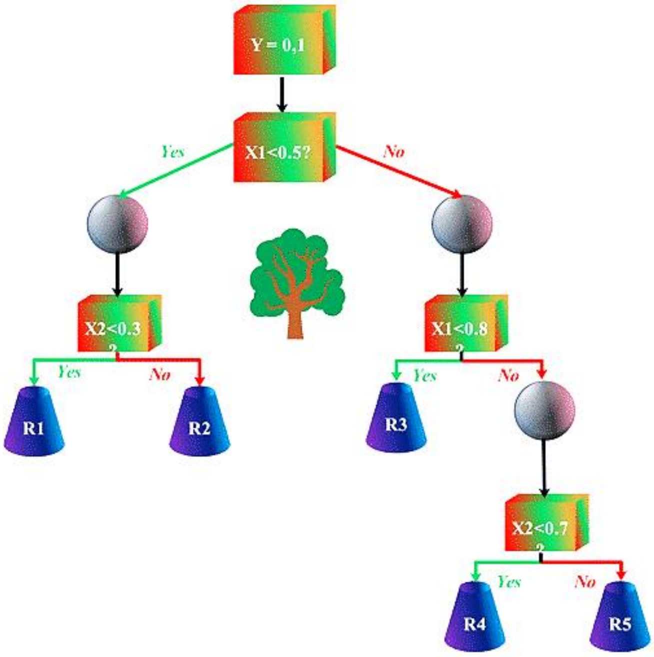

Figure 1. Graphical representation of basic decision tree.

Nodes and branches form the essential elements of a DT model, with the key steps in constructing the model involving the processes of splitting, stopping, and pruning.

2.2. Random Forest Regression (RFR)

The RF approach produces a large number of separate decision trees that function as regression models. The mean of these decision trees determines the final result [10]. Each decision tree in the Classification and Regression Tree (CART) learning process consists of decision nodes and leaf nodes depending on the input vector X and scalar output Y. The learning process in CART takes into account the substantial impact of input data complexity on tree development [11]. Equation (1) provides a mathematical expression for the set of n observations in the train set, which is referred to as R_n

$$R_n = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}, X \in R^m, Y \in R \tag{1}$$

In the train stage, the algorithm refines split function parameters to partition input data at each node, aligning with the R_n set. The initial use of a decision tree starts at the primary node, iteratively dividing the dataset until terminal nodes or leaves are reached. Arboriculture often involves restricting tree growth based on predetermined maximum levels or when a node has observations below a specified threshold, with the outcome prediction carried out by the generated prognostic function $\widehat{H} = (X, R_n)$ after the training process.

In the realm of RFR, a set of L tree-structured base models denoted as H=(X, Θ_K), where K takes values from 1 to L and Θ_K consists of independent and identically distributed random vectors, is applied. The process of constructing a RF [11] involves the random selection of either a portion of the training dataset or a portion of the characteristics for every DT.

The bootstrap method involves randomly selecting a sample by selecting n observations from Rn with repetitive selection, each having an equal probability of 1⁄n The bagging algorithm selects multiple bootstrap samples $(S_n^{\Theta_1}, \ldots, S_n^{\Theta_q})$ exposed to the tree decision algorithm, resulting in a collection of q prognostication trees denoted as $\hat{h}(X, S_n^{\Theta_1}), \ldots, \hat{h}\left(X, S_n^{\Theta_q}\right)$. The ensemble produces output values $\widehat{Y}_1$, $\widehat{Y}_2, \ldots, \widehat{Y}_q$ representing predictions by individual trees, aggregated by calculating the results' average value, leading to the predicted value of Y($\widehat{Y}$) as described in [40].

$$\widehat{Y} = \frac{1}{q}\sum_{l=1}^{q} \widehat{Y}_l = \frac{1}{q}\sum_{l=1}^{q} \widehat{H}(X, R_n^{\Theta_l}) \tag{2}$$

$\widehat{Y}_l$ is the output of *l-th* tree, and *l*=…1, 2, , *q*. During training, certain data may be utilized repeatedly while others go unused, potentially influencing the general effectiveness of the learning algorithm. By incorporating the bagging approach into RF modeling, the model's stability is increased and the RF regression algorithm's resistance to small dataset inconsistencies is strengthened. Interestingly, individual tree development happens without pruning, resulting in a lightweight and computationally efficient model.

The RF regression technique Figure 2, highlighted in [29], is characterized by its simplicity, necessitating adjustments to two parameters: the number of trees (n_{tree}) and the randomly-selected attributes for each forest partition (m_{try}) Augmenting tree density in a forest has the potential to improve resilience and precision in prognostic models, yet it comes with an increased computational load. Despite the observed convergence of generalization error with a higher number of trees, the normal value of n_{tree}=500 is frequently utilized, highlighting the individual potency of each decision tree and the potential inter-tree dependence within the forest [36].

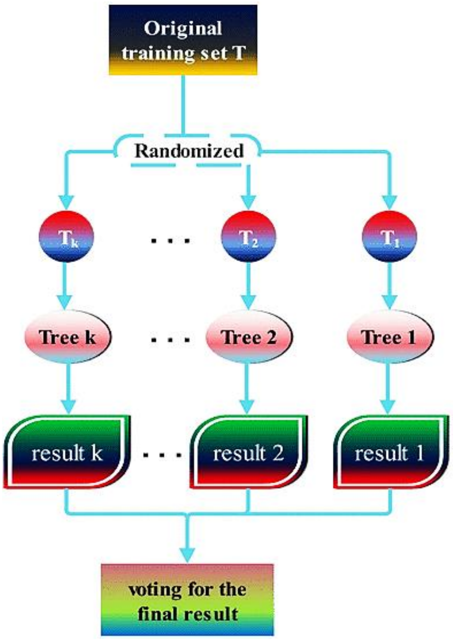

Figure 2. Graphical representation of RFR.

The generation of n_{tree} bootstrap examples involve randomly selecting subsets from the original dataset, each with around two-thirds of the elements. For unpruned regression trees, a modification includes a randomized approach in choosing predictors for node splits by selecting a subset of m_{try} predictors. Combining predictions from the n_{tree} sub-models through a majority voting scheme for classification and averaging for regression enhance the accuracy and reliability of the overall predictive model, especially for anticipating novel data.

By picking data at random from the original dataset with replacement, a new training set of bootstrap samples is formed, which is then used to construct a regression tree. It is not necessary to use every sample set while bagging a tree; some data may be used more than once. Additionally, the performance of the regression tree can be assessed by including unselected data into Out-Of-Bag (OOB) samples. The likelihood of overfitting may be considerably decreased by using RF.

2.3. Rhizostoma Optimization Algorithm (ROA)

The ROA is inspired by the Rhizostoma octopus, also known as the dustbin-lid jellyfish or frilly-mouthed jellyfish, celebrated for its colossal size [31]. The algorithm is designed based on three fundamental principles: firstly, emulating Rhizostoma pulmo's exploration for optimal food locations in the ocean using a combination of random searching algorithms, forming a swarm with exploration and feeding motions; secondly, incorporating two main bands of a meta-heuristic algorithm exploration as the first type of motion and exploitation as the second type of motion within the swarm; and thirdly, implementing a motion control factor to govern the switch between searching for food and moving within the swarm, while evaluating the quality of found food based on the current location and its associated objective function.

2.3.1. Setting up the Population and Defining Boundary Conditions

The R octopus population is initialized randomly, as formulated in Equation (3):

$$X_i^* = \lambda x_i(1 - x_i), 0 \le x_0 \le 1 \quad (3)$$

R octopus' locations, denoted by x_i, are randomlyinitialized within the range [0, 1] based on the initial population x_0, comprising values {0, 0.25, 0.5, 0.75, 1}. In the event of surpassing search area boundaries, a R. octopus re-enters from the opposite bound, as specified in Equation (4):

$$X_{i,d}^* = \begin{cases} (x_{i,d} - Ub_d) + Lb_d, if\ x_{i,d} > Ub_d \\ (x_{i,d} - Lb_d) + Ub_d, if\ x_{i,d} < Lb_d \end{cases} \quad (4)$$

Representing the R. octopus' location in the d_{th} dimension, $x_{(i,d)}$ transforms to $x_{i,d}^*$, considering boundary constraints, where Ub_d and Lb_d signify the upper and lower bounds in the search spaces.

2.3.2. Strategies for Locating Food

Changing movement patterns of R. octopus over time prompted the proposal of Fast Simulated Annealing (FSA) to model its behavior [1]. Initially, it was believed that R. octopus employed Levy Walks (LW) similar to sharks and honeybees. These foraging strategies involve a generative function guiding variable updates during each search attempt. Here, the ROA was introduced, employing LW, FSA, and SA to determine the most effective motion strategy for R. octopus. Levy flight, named after mathematician Paul Levy, involves random walks with step lengths following a Fibonacci distribution. If the fitness of a potential new location surpasses the current location, R. octopus moves towards it, following a heavy-tailed Levy distribution. Equation (5) represents this distribution's power-law behavior.

$$L(s) \approx |s|^{1-\beta}, where\ 0 < \beta \le 2 \quad (5)$$

The step length s in the random Levy distribution is determined by the formula $s=u/|v|^{(1/\beta)}$, where u and v are derived from normal distributions. The update of the position, giving maximum consideration to food, can be represented as expressed in Equation (6):

$$x_{next} = x_i + \mu \times rand(0,1) \quad (6)$$

The determination of the step length (μ) involves the multiplication of a uniformly distributed random number (rand) between 0 and 1. In the FSA, a cost function compares prey density at the current position with another position, and the step length s is randomly drawn from a Cauchy distribution as |y-x|.

$$p(s) = 1/\pi \times T/(s^2 + T^2) \tag{7}$$

T represents the temperature, serving as a measure of the extent of fluctuations in step length. The calculated probability value needed to accept the new position is expressed in Equation (8):

$$p = min\{1, exp(\Delta f/T)\} \tag{8}$$

The change in cost, Δf=*f*(*y*)-*f*(*x*), determines the difference between the current and previous positions. Acceptance of a new position depends on improvement (Δf >0); otherwise, the R. octopus reverts to its previous location. FSA converges to the global optimum with annealing scheduled as $T(k)=T_0/k$, and another strategy involves Simulated Annealing (SA), a variant of Levy flight with step lengths randomly generated from a Gaussian distribution.

$$g(s) = (2\pi T)^{-D/2}{}_{e^{(\Delta x^2)}/2T} \tag{9}$$

D represents the dimension of the search space, which corresponds to the number of variables in the cost function. The rate of change of the variables' vector is denoted by Δx, and the transition from the current state x_i to the next stage of variables is expressed as $x_{next}=x_i+\Delta x$.

2.3.3. Swarm of R. Octopus

In the Rhizostoma swarm, there are two distinct waves: feeding motion and forming swarm. Initially, the swarm exhibits feeding motion, and as time progresses, it transitions to forming swarm motion. The first type involves the movement of R. octopus around their respective locations, with the updated position determined by Equation (10):

$$x_i(t+1) = x_i(t) + \beta \times r\ and(0.1) \times (ub - lb) \tag{10}$$

The search space is defined by *ub* and *lb* bounds, with a motion coefficient β >0 Optimal results in the ROA are observed when β.is set to 0.1. The second type of motion involves the random selection of R_j and R_i, with movement directed by a vector from R_i to R_j R. octopus moves towards locations with more food and away from those with less, fostering swarm formation. Equation (11) simulates the updated location and motion direction of a R. octopus.

$$\vec{S} = x_i^{(t+1)} - x_i^t \tag{11}$$

$$\vec{S} = rand(0,1) \times \vec{D} \tag{12}$$

$$\vec{S} = \begin{cases} x_j^t - x_i^t\ , & if\ f(x_i) \geq f(x_j) \\ x_i^t - x_j^t\ , & if\ f(x_i) < f(x_j) \end{cases} \tag{13}$$

f represents the impartial function for place *x*, $\vec{S}$ denotes the step and $\vec{D}$ indicates the direction.

$$x_i^{(t+1)} = x_i^t + \vec{S} \tag{14}$$

2.3.4. Factor Regulating Motion

A wave control feature dictates the type of motion, overseeing both swarm behaviors and R. octopus' food-search strategies. Rhizostoma pulmo is drawn to locations abundant in plankton, leading to swarm formation. The factor, represented by the function *M*(*f*), randomly varies from 0 to 1. If *M*(*f*) is less than the constant m_0 (set at 0.5), R. octopus adopts an individual food search strategy; otherwise, individual octopuses engage in swarm movement.

$$M(f) = |1 - exp((t-1)/t_{max})(2 \times rand(0,1) - 1)| \tag{15}$$

t represents the current iteration number, and t_{max} is the initialized parameter indicating the max number of iterations. The pseudocode for ROA is represented below. The adaptive values of *M*(*f*) facilitate an exploration/exploitation balance, enabling a smooth transition between the two processes and an equal distribution of iterations.

```
Begin
Initiate the parameters, encompassing the population magnitude (n), the upper limit for iterations (T), and the assembly of R octopuses.
    $x_i = (1,2,\dots,n)$,
Determine the performance score of each exploration agent through the evaluation function f(x).
$X^*$ represents the optimal current position.
Commence iteration with t set to 1.
Execute the process again.
    For i=1: $n_{pop}$.do
        Derive the movement switch aspect, M(f), through the application of the formula provided in Equation (15).
        Else: The R. octopus navigates within a collective of its peers.
            If $r\ and(0,1) > (1 - M(f))$: The initial category of movement is displayed by R. octopus.
                Adjust the location of the existing search agent utilizing the equation provided in Equation (10)
            Else: R. octopus demonstrates a different form of motion.
                Ascertain the orientation of R. octopus using the equivalence outlined in Equation (13)
                Revise the location of the existing search agent based on the formula in Equation (14)
            End If
        End If
        Monitor and adjust search agents to stay within defined boundaries.
        Determine the health or effectiveness of each search agent through a fitness calculation.
        Revise $X^*$ in the presence of an improved solution.
    End for i
    Increment the iteration count by updating t to t+1.
Continue the process until the specified stopping condition is satisfied (t>t_max)
Continue until the stopping condition is satisfied.
Present the optimal outcome along with graphical representation.
End
```

2.4. Dwarf Mongoose Optimizer (DMO)

The DMO algorithm, inspired by dwarf mongooses' foraging behaviors, features a stochastic population-based approach. Individual mongooses conduct independent food searches, while semi-nomadic tendencies influence the collective foraging process [1], constructing resting mounds near abundant food sources. The program solves optimization issues by simulating the mongoose lifestyle via mathematical modeling [3, 42].

The DMO algorithm starts by probabilistically generating a population of candidates within predefined lower and upper bounds for problem-solving.

$$k = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d-1} & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d-1} & x_{2,d} \\ \vdots & \vdots & x_{1,1} & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,d-1} & x_{n,d} \end{bmatrix} \tag{16}$$

X denotes the existing population of candidates, generated randomly through Equation (17), with each element $x(_{i,j})$ representing the location of the j_{th} dimension inside the i_{th} element. Significantly, n represents the number of elements in the population, while d specifies the dimensionality of the problem at hand.

$$x_{i,j} = unifrnd(VarMin, VarMax, VarSize) \tag{17}$$

The DMO algorithm uses the variable *unifrnd* as a consistently distributed chance number, *VarMin* and *VarMax* representing problem boundaries, and *VarSize* indicating the size of the difficulty's dimensions. The method consists of two steps that follow a typical metaheuristic approach: exploitation, involving an intensive search (intensification) within designated spaces, and exploration, encompassing a randomized search for novel resources (diversification). The alpha group, the scout group, and the babysitters are the three primary social institutions that carry out these tasks.

2.4.1. Alpha Group

The alpha female (α) is chosen using Equation (18), giving her the power to guide the family.

$$\alpha = \frac{fit_i}{\sum_{i=1}^{n} fit_i} \tag{18}$$

The number of mongooses in the alpha group is determined by the value of n-bs, bs represents the individuals responsible for nursing and calm for young children. Additionally, the abundance of nutrients positively affects the construction of the resting mound, as shown by Equation (19):

$$X_{i+1} = X_i + \varphi * peep \tag{19}$$

The variable φ is a uniformly distributed number within the range [-1, 1]. During each iteration, the evaluation of the sleeping mound occurs, as expressed in Equation (20):

$$sm_i = \frac{fit_{i+1} - fit_i}{max\{|fit_{i+1}, fit_i|\}} \tag{20}$$

When a latent accumulation is discovered, Equation (21) is used to calculate a mean numerical value:

$$\rho = \frac{\sum_{i=1}^{n} sm_i}{n} \tag{21}$$

2.4.2. Scout Group

After meeting the babysitter argument criteria, the subsequent phase includes scouting to identify a fresh sleeping site linked to a specific food source. Observing the mongoose's tendency to avoid returning to previous sleeping mounds, the scouting team seeks a new location. The mongoose displays a typical conduct of simultaneously searching and scouting in a DMO, where Equation (22) illustrates the process by emphasizing increased distance for a higher likelihood of discovering the next sleeping mound.

$$X_{i+1} = \begin{cases} X_i - CF * phi * rand * [X_i - \vec{M}] \ if \ \rho_{i+1} > \rho_i \\ X_i + CF * phi * rand * [X_i - \vec{M}] \qquad else \end{cases} \tag{22}$$

$$CF = \left(1 - \frac{iter}{Max_{iter}}\right)^{\left(2\frac{iter}{Max_{iter}}\right)} \tag{23}$$

$$\vec{M} = \sum_{i=1}^{n} \frac{X_i \times sm_i}{X_i} \tag{24}$$

R and generates a random number within [-1, 1], while the CF parameter influences the collective behavior of mongooses by linearly decreasing over iterations and the vector $\vec{M}$ drives the displacement of mongooses towards a new sleeping mound.

2.4.3. Babysitters Group

When group members postpone foraging or scouting until they reach the babysitting exchange parameter in Equation (22) the candidate population decreases, and the caretaker cohort takes care of the juveniles while the scouting unit looks for a place to rest and food.

The outlined algorithm is represented in the following pseudocode:

```
Stablish the Algorithm's Parameters:

Produce
For $iter = 1: max\_iter$
Calculate the fitness of the mongoose
Initialize the time counter C
Set time counter
Initiate the time-tracking counter.
Calculate alpha using the formula in Eq.(18).
Calculate a potential food position using the formula in Eq.(19).
Evaluation of novel fitness of X_(i+1)
Guesstimate the asleep using Eq.(20).
Compute the average value of the sleeping mound using the formula in Eq.(21).
Compute the movement vector using.
Determine the movement vector through computation using Eq.(24).
Conversation babysitters if C ≥ L.
Exchange babysitters when C equals or surpasses L.
Establish the position of bs and calculate its fitness
$fit_i \leq \alpha$
Compute the prospective position of the scouting mongoose using the formula in Eq.(22).
Update the best solution achieved thus far.
```

```
End For
Return the greatest answer
End
```

3. Dataset Preparation

3.1. Data Engineering

This section outlines the process of selecting, engineering, and cross-validating data for ML models in predicting the market value of football players. Then, justifies the choices made regarding feature selection or exclusion.

The dataset utilized in this research is sourced from the FIFA 19 video game (as a simulator developed by EA Sports), along with real-world statistical reports (https://www.openml.org/search?type=data&status=active&id=43604). Popularity of FIFA games provides an interesting parallel with real-world statistical reports in the field of football. Real-world statistics, including player performances, team dynamics, and strategic analyses, complement the virtual experience by offering insights into the broader football landscape. This broad dataset initially contained 53 features for 491 sample players which needed data engineering to become appropriate for market value estimation of well-known football leagues' players.

During the initial stage of dataset cleaning, seven samples were excluded from the dataset because of the lack of some feature values. Then, the corresponding leagues of each player are extracted according to the clubs' name, and a new column is added to the dataset as 'league name.' A dominant number of 459 sample players were related to Serie A, Premier League, League 1, La Liga, and Bundesliga, and the remaining 25 sample players from less professional leagues with less valuable players (outliers) were eliminated from the dataset. Finally, some feature columns, such as players' names and nationality with no analytical purpose removed, and some other nominal features, such as preferred foot, weak foot, and league name, were labeled to be appropriate for ML regression tasks.

Table 1. Statistical analysis of market values of footballers in each European league.

League	Number of players in the dataset	Statistical evaluators of market values (dollars)		
		Maximum	Average	Minimum
Serie A	103	89000000	16530097	325000
Premier League	97	93000000	23804124	2700000
League 1	85	118500000	12731176	400000
La Liga	82	110500000	21202439	1200000
Bundesliga	92	77000000	15017935	750000

The number of players corresponding to each league and the Minimum, maximum, and average values of players in the transfer market of specified leagues are reported in Table 1 to give insights about the range of salaries clubs of leagues allocate for players for further discussion in the following sections.

3.2. Dataset Cross-Validation

The exponential growth of complex datasets demands innovative techniques for data extraction, as traditional methods fall short [50]. As the large dataset selected in this study contains sample players from five different European leagues, the cross-validation procedure guarantees the reliable performance of predictive models in dealing with smaller datasets. The k-fold approach is an efficient form of cross-validation that examines the generalization ability of the model by separating the samples in the dataset into a specified number of folds [45]. Then, the learning process is iterated K times, and in each iteration, one of the folds is used as a training dataset. In this study, K set to be ten, and as it is illustrated in Figure 3, in each iteration, 10% of the samples used for train, and the remaining 90% used for validation. R^2 measures are documented for each iteration (Figure 4). The analysis of the model results revealed that, for the studied dataset, the best outcome was achieved with the highest average R^2 value equal to 0.976 in K9.

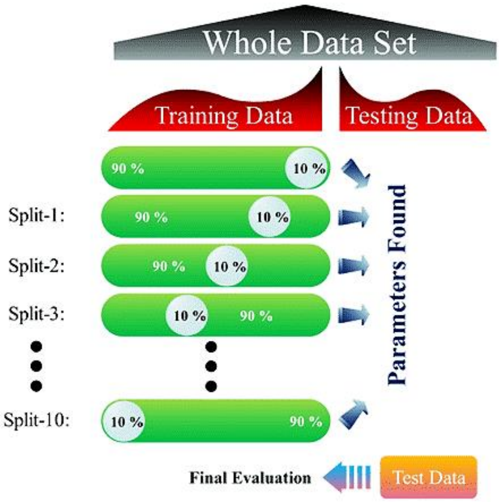

Figure 3. Schematic description of 10-fold cross-validation.

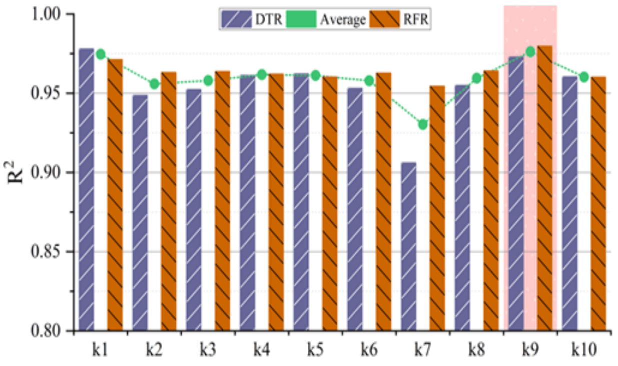

Figure 4. Representative R2 values for prediction performance of DTR and RFR models in each iteration of cross-validation.

3.3. Selecting Informative Features

After data preprocessing, there were 46 variables in the model to be used for predicting the market values of players of five leagues. Due to the interdependence of various statistics, such as goals a component of total shots on goal, and the correlation between games played and minutes played, it was imperative to examine the dataset for potential multicollinearity issues. The Variance Inflation Factor (VIF) identifies variables demonstrating substantial correlations with one or more other variables within the dataset. High VIF scores suggest a greater presence of explanatory variables contributing to multicollinearity concerns [20].

$$VIF = \frac{1}{1 - R_i^2} \tag{25}$$

R_i^2 is the coefficient of determination for *ith* input feature in the dataset. According to Table 2, within the dataset of this study VIF values considering the initial feature vector were in the range of 16.26 to 1.15. A strategy employed to address multicollinearity involves systematically eliminating variables based on their VIF scores. This method entails iteratively removing variables, starting with the one exhibiting the highest VIF score. The removal of a variable subsequently reduces the VIF scores of other variables [26]. Conventionally, an acceptable VIF score falls within the range of 5 to 10, which is established to be 5 in this study as the initial VIF score was not high values [4]. An iterative procedure was implemented, wherein the function successively eliminates the variable with the highest VIF score, recalculates scores for the remaining variables, and continues this process until all remaining variables attain VIF scores below the specified threshold of 5. After feature vector reduction, 33 variables are left, for which the correlation matrix in Figure 5 illustrates the relationship between these factors. Some selected features of a representative player from each league are presented in Figure 6.

Table 2. Description of features and determination of selected features.

No.	Feature	Description	VIF score-initial feature vector	VIF score-reduced feature vector
1	Age	The player's age (affecting both the experience and future potential)	2.03	1.94
2	League label	Playing league of footballer(Serie A, Premier League, League 1, La Liga, and Bundesliga)	1.26	1.20
3	Preferred foot	Primary or adept foot utilized for shooting, passing, and dribbling	1.27	1.22
4	International reputation	Globally recognized reputation (estimated through fans' comments in social network applications)	2.99	2.83
5	Weak foot	The player's inability to effectively use both legs in football.	1.24	1.21
6	Skill moves	Techniques performed to outmaneuver opponents involve intricate ball control, dribbling, and feints	2.90	2.47
7	Height	Height of the player (affecting the likelihood of scoring or preventing a goal).	4.73	4.42
8	Weight	Weight of the player (affecting the movement skills of the players)	3.47	3.34
9	Crossing	Player's technique in delivering the ball into the penalty area from the flanks.	4.39	4.10
10	Finishing	Player's ability to successfully score goals	10.09	Removed
11	Heading accuracy	The player's proficiency in heading the ball.	3.99	3.72
12	Short passing	The number of passes to teammates and the accuracy of passing	7.64	Removed
13	Volleys	Striking the ball while it is in the air	5.41	4.50
14	Dribbling	controlled touches to maneuver the ball while on the move	10.30	Removed
15	Curve	Bending or swerving the ball during a shot or a pass	4.87	4.62
16	Free kick accuracy	Accuracy of free kicks.	3.90	3.57
17	Long passing	The number of long passes to a teammate.	6.51	3.63
18	Ball control	Skilfully receiving, trapping, and manipulating the ball	9.30	Removed
19	Acceleration	How quickly a player can reach their top speed	11.36	Removed
20	Sprint speed	Maximum velocity of a player during a full-out sprint	8.78	2.44
21	Agility	Rapidly direction change by player	6.16	Removed
22	Reactions	Quick responses to the movement of the ball and the actions of opponents and teammates	4.10	3.73
23	Balance	Ability to maintain stability during various movements	5.44	4.59
24	Shot power	strength of strikes on the ball	4.12	2.88
25	Jumping	Jumping ability of the player	2.10	1.97
26	Stamina	Ability to sustain physical effort and performance over an extended period of playing	2.43	2.28
27	Strength	Physical power and ability to exert force against resistance	4.38	4.10
28	Long shots	Successful shots from a considerable distance away from the goal	5.80	Removed
29	Aggression	Assertiveness in challenging for the ball	2.86	2.12
30	Interception	Successfully blocks a pass or a ball played by the opposing team	9.29	Removed
31	Positioning	Detected positions for players during games	7.68	4.17
32	Vision	Ability to perceive the unfolding dynamics of the game	7.43	Removed
33	Penalties	Penalty kicks by a player	2.75	2.59
34	Composure	Ability to maintain calmness and focus in high-pressure situations during a match	4.33	3.87
35	Marking	Tracking and guarding an opponent to prevent them from receiving the ball	4.13	2.83
36	Standing tackle	Number of standing tackles	16.26	Removed
37	Sliding tackle	Number of sliding tackles	12.65	Removed
38	Games played	Number of matches played	3.69	3.25
39	Games started	Number of matches started by the player	10.04	Removed
40	Minutes Played	Overall playing time (minutes)	9.36	4.20
41	Goals	The number of goals scored	4.98	2.98
42	Assist	Assisting other players in scoring a goal	1.94	1.88
43	Shots on goal	Number of shots of a player toward the goal	15.36	Removed
44	Shots	Total number of shots by player	11.77	4.33
45	Yellow card	Frequency of yellow cards received	1.55	1.49
46	Red card	Frequency of Red cards received	1.15	1.13

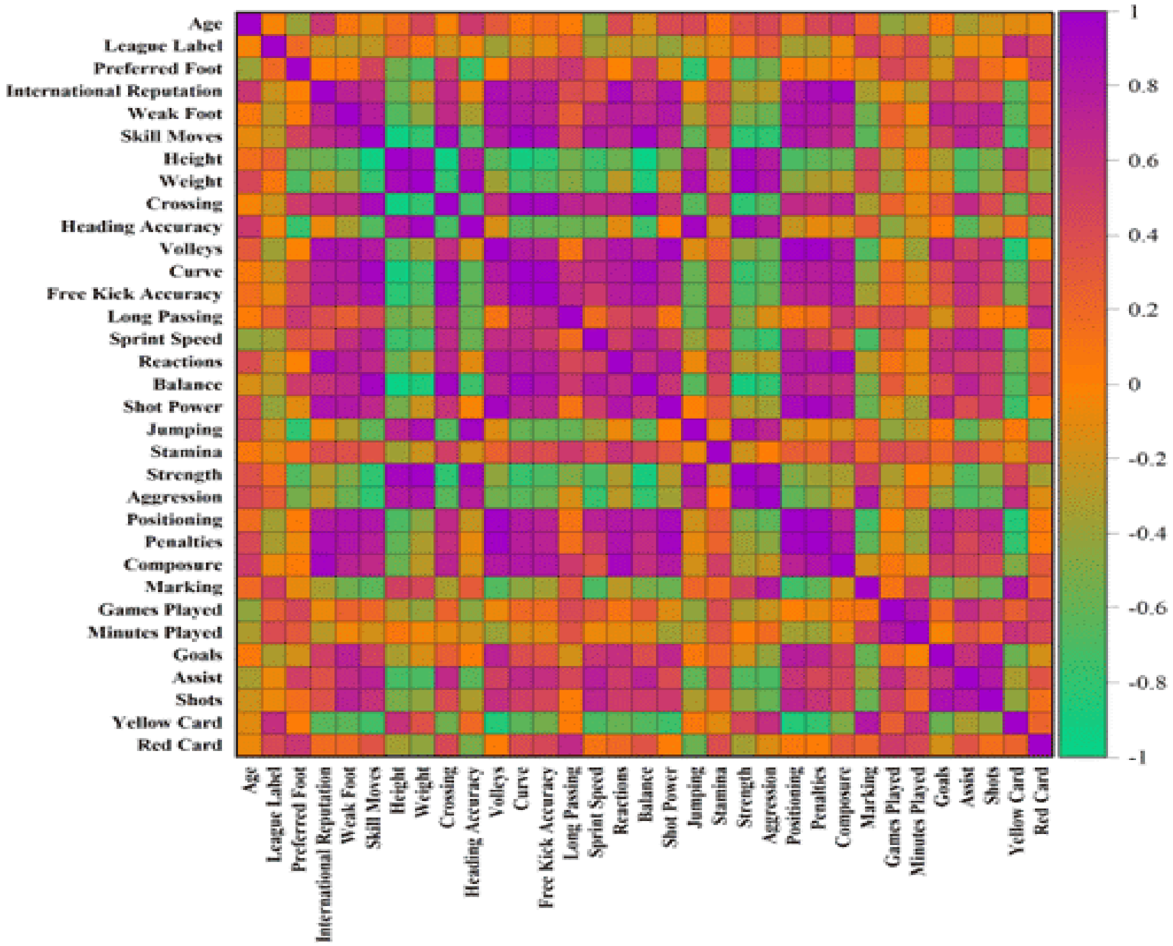

Figure 5. Correlation between selected features as input variables for ML models' training.

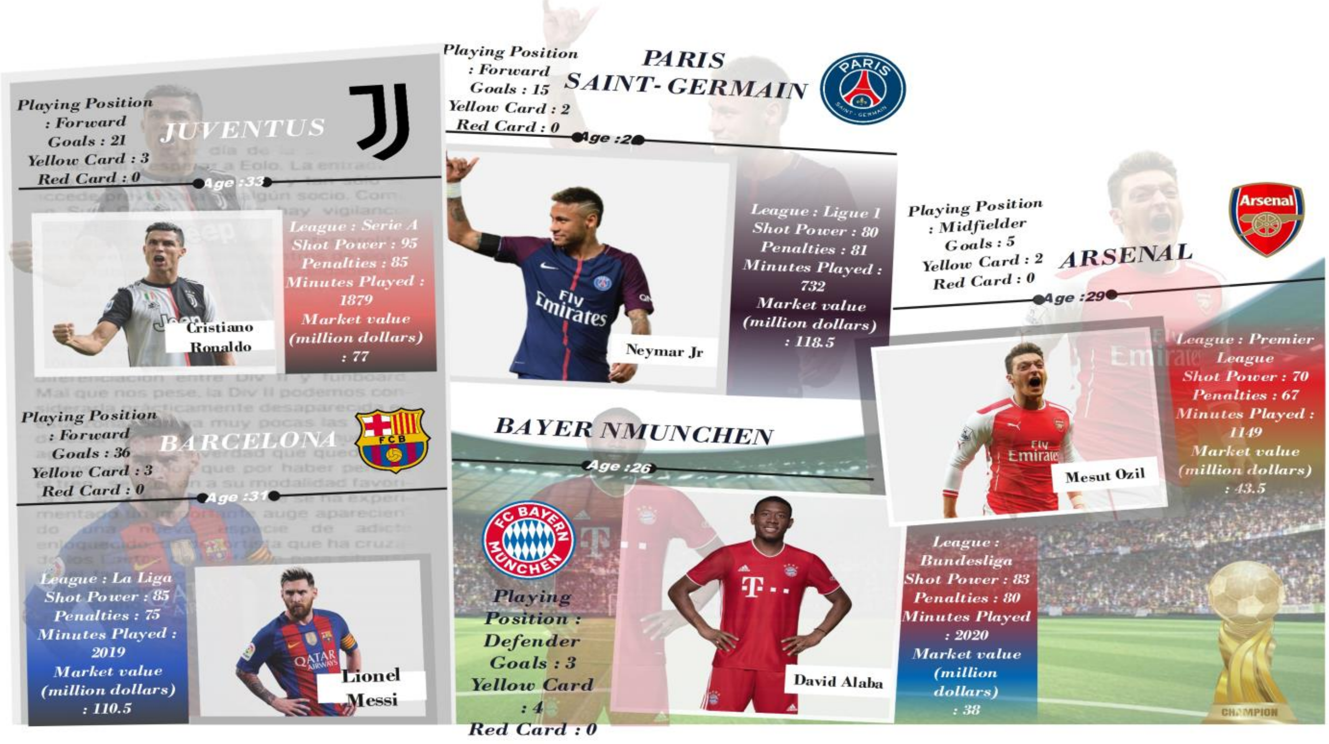

Figure 6. Important features of representative players from each European league.

Finally, it is worthwhile to note that for the prediction task, the dataset with 459 sample players, which contained 33 features for each player randomized to prevent any inherent order or pattern in the data from influencing the learning process. Then, the samples in the dataset were divided into 70%, 15%, and 15% proportions for training, validating, and testing the proposed prediction models.

4. Prediction Results

4.1. Performance Evaluation and Validation Metrics

Assessing the accuracy of predictions by developed models is challenging due to the unobservability of market values. While market values are proxies for transfer fees, a direct comparison was made with actual transfer fees, acknowledging their inherent differences [22]. The following metrics are powerful tools for evaluating the prediction accuracy of developed ML models:

- **Coefficient of Determination (R2)**

Ranging from 0 to 1, the coefficient of determination expresses how much of the variation in the dependent variable can be attributed to the independent factors. An R² of 1 specifies perfect prediction, while 0 signifies the model’s inability to explain any variance.

$$R^2 = \left(\frac{\sum_{i=1}^{n}(M_i - \bar{M})(P_i - \bar{P})}{\sqrt{\left[\sum_{i=1}^{n}(M_i - \bar{P})^2\right]\left[\sum_{i=1}^{n}(P_i - \bar{P})^2\right]}}\right)^2 \tag{26}$$

- **Root Mean Square Error (RMSE)**

RMSE is a quantitative measure of the accuracy of a model’s predictions. Its utility lies in providing a measure of the dispersion of residuals, indicating how well the model aligns with the actual data points.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(M_i - P_i)^2}{n}} \tag{27}$$

- **Mean Square Error (MSE)**

MSE measures the average of the squared differences between predicted and observed values in regression models. By emphasizing larger errors due to the squaring process, MSE provides a comprehensive assessment of a model’s accuracy and precision.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(M_i - P_i)^2 \tag{28}$$

- **Mean Absolute Relative Error (MARE)**

MARE estimates the mean of the absolute relative errors, which are the absolute differences between predicted and observed values divided by the observed values.

$$MARE = \frac{1}{n}\sum_{i=1}^{n}\frac{|M_i - P_i|}{|\bar{M} - \bar{P}|} \tag{29}$$

- **Nash-Sutcliffe Efficiency (NSE)**

The NSE assesses the model’s performance by comparing the squared differences between actual and estimated values to those between observed values and their mean. With a range from negative infinity to 1, a perfect match is indicated by 1, while values below zero suggest that using the mean of observed data would be a better predictor than the model.

$$NSE = 1 - \frac{\sum_{i=1}^{N}(P_i - M_i)^2}{\sum_{i=1}^{N}(M_i - \bar{M})^2} \tag{30}$$

where M_i and $\bar{M}$ are the measured and average measured values, P_i and $\bar{P}$ are the predicted and average predicted values, and n is the total number of data.

- **Optimization of Machine Learning Models**

DMO and ROA were two recently developed optimization algorithms utilized in this study for hybrid models development. Optimized hyperparameters of RFR and DTR models by each of the optimizers are reported in Tables 3 and 4. The convergence trend of the hybrid models is depicted in Figure 7, revealing that DTR-based models exhibited error values twice as high as RFR-based counterparts at the initial point of 200 optimization iterations. RFRO, as the optimal model, had a minimum error in the start point of operation, and during around 130 iterations, ROA decreased error values of RFR by more than four-fold (final error of 2 million dollars).

Table 3. Optimized hyperparameters of DTR.

Hyperparameter	Models	
	DTDM	**XGNG**
max_depth	36	999
min$_{samples}$ *split*	0.000172	0.001
min$_{samples}$*_leaf*	0.002604	0.0005
max$_{leaf}$ *nods*	120	3310

Table 4. Optimized hyperparameters of RFR.

Hyperparameter	Models	
	RFDM	**RFDM**
n_estimators	16	53
max_depth	7	327
min$_{samples}$ split	2	2

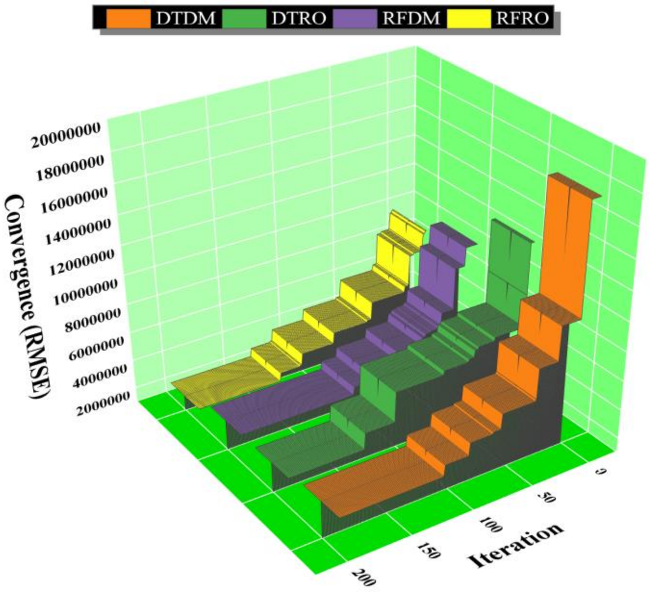

Figure 7. Convergence pattern of hybrid models in market value prediction.

To improve the reliability of approximations, investigating the optimization capability of various algorithms such as DMO and ROA and their combinations is a novel approach. This combination (ensemble) integrates outputs from individual models for a unified mathematical expression.

$$Y_{ensemble} = \sum_{i=1}^{n} \frac{Y_{model_i}}{n} \tag{31}$$

$Y_{ensemble}$ signifies the output of an ensemble, Y_{model_i} denotes the output produced by the *i-th* model, and n represents the total number of algorithms incorporated within the ensembling.

Moreover, optimizers may exhibit varying levels of enhancement capabilities. Consequently, it becomes necessary to modify the impact of each optimizer within the ensemble according to its performance. This methodology is referred to as a weighted averaging ensemble, and the ensemble's prediction $Y(x)$ expressed as:

$$Y(x) = \sum_{i=1}^{n} \omega_i \times y_i(x) \tag{32}$$

$$\sum_{i=1}^{n} \omega_i = 1 \tag{33}$$

Where $y_i\,(x)$ represents the prediction of the *i-th* model within the ensemble, ω_i denotes the corresponding weights assigned to the *i-th* model, and n is the total number of models.

4.2. Metric Results

In the analysis of the metric results reported in Table 5, it is apparent that R2 values serve as crucial indicators of predictive performance. Notably, the consistently higher R2 values observed in RFR-based models compared to DTR-based ones suggest that RFR models generally outperform DTR models in explaining the variance in football player market values. The exceptionally high R2 value of approximately 98% for ensemble models underscores the efficacy of combining multiple models for improved predictive accuracy. The interpretation of prediction errors provides further insights into the models' performance. For instance, the minimum error of 2 million dollars achieved by RFRO, amounting to approximately 12% of the mean market values across all leagues, suggests a high level of precision in predicting player values. Conversely, the higher prediction errors observed in DTRO and RFDM models, around 3.1 million dollars each, imply a greater degree of uncertainty and lower accuracy in their predictions. These nuanced interpretations of metrics highlight the strengths and weaknesses of different models, providing valuable guidance for decision-makers in the football industry seeking to optimize player selection and budget allocation strategies based on historical performance data as shown in Figure 8.

Table 5. The result of developed models for DT and RFR.

Model	Category	Phase	Index values				
			RMSE	R2	MSE	MARE	NSE
RFR	Single	Train	3.1E+06	0.9744	9.77E+12	1.64E-01	9.70E-01
		Validation	4.5E+06	0.9426	2.00E+13	2.17E-01	9.35E-01
		Test	4.5E+06	0.9623	2.01E+13	2.75E-01	9.43E-01
		Total	3.6E+06	0.9671	1.29E+13	1.88E-01	9.61E-01
RFDM	Hybrid	Train	2.7E+06	0.9827	7.36E+12	2.24E-01	9.77E-01
		Validation	3.5E+06	0.9697	1.21E+13	2.38E-01	9.61E-01
		Test	4.0E+06	0.9747	1.64E+13	3.01E-01	9.53E-01
		Total	3.1E+06	0.9787	9.43E+12	9.43E+12	9.43E+12
RFRO	Hybrid	Train	1.7E+06	0.9931	2.98E+12	9.03E-02	9.91E-01
		Validation	2.5E+06	0.9838	6.42E+12	1.30E-01	9.79E-01
		Test	2.8E+06	0.9879	7.79E+12	1.71E-01	9.78E-01
		Total	2.0E+06	0.9905	4.22E+12	1.08E-01	9.87E-01
RFRD	Ensemble	Train	2.1E+06	0.9896	4.57E+12	1.52E-01	1.52E-01
		Validation	2.9E+06	0.9791	8.59E+12	1.77E-01	9.72E-01
		Test	3.4E+06	0.9834	1.13E+13	2.33E-01	9.68E-01
		Total	2.5E+06	0.9865	6.18E+12	1.68E-01	9.81E-01
DTR	Single	Train	4.2E+06	0.9478	1.79E+13	3.98E-01	9.45E-01
		Validation	5.2E+06	0.9160	2.74E+13	2.36E-01	9.11E-01
		Test	4.5E+06	0.9429	2.05E+13	3.25E-01	9.42E-01
		Total	4.4E+06	0.9407	1.97E+13	3.63E-01	9.40E-01
DTDM	Hybrid	Train	3.3E+06	0.9668	1.11E+13	1.53E-01	9.66E-01
		Validation	4.7E+06	0.9279	2.23E+13	2.40E-01	9.28E-01
		Test	3.2E+06	0.9752	1.04E+13	2.35E-01	2.35E-01
		Total	3.6E+06	0.9624	1.27E+13	1.78E-01	9.61E-01
DTRO	Hybrid	Train	2.9E+06	0.9750	8.33E+12	2.72E-01	9.75E-01
		Validation	3.0E+06	0.9744	8.82E+12	1.50E-01	9.71E-01
		Test	3.9E+06	0.9701	1.50E+13	2.38E-01	9.57E-01
		Total	3.1E+06	0.9715	9.41E+12	2.49E-01	9.71E-01
DTRD	Ensemble	Train	2.2E+06	0.9858	4.74E+12	1.60E-01	9.85E-01
		Validation	2.8E+06	0.9743	7.91E+12	1.70E-01	9.74E-01
		Test	3.0E+06	0.9838	9.09E+12	2.05E-01	9.74E-01
		Total	2.4E+06	0.9829	5.87E+12	1.68E-01	9.82E-01

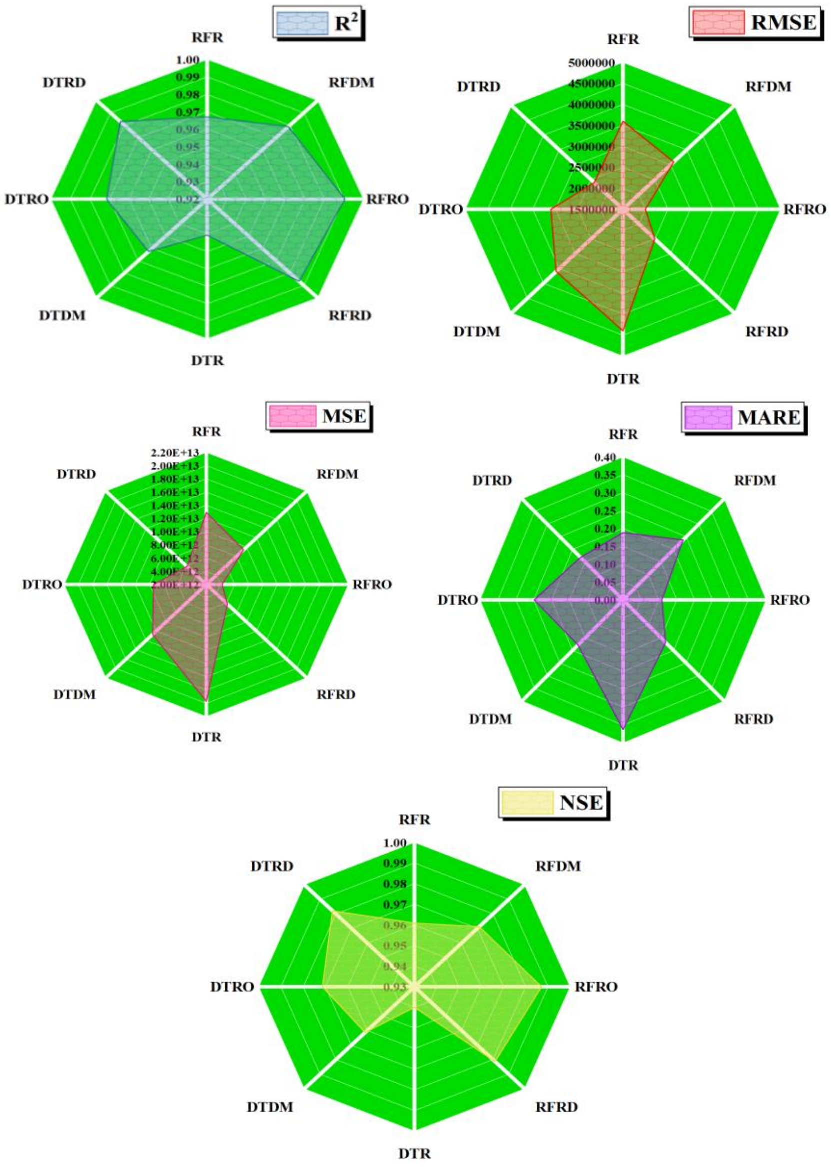

Figure 8. Comparative representation of prediction performance metric results.

5. Discussion and Future Works

5.1. Comparative Analysis of Predictors' Performance

Visual representation of predictors' performance by scatter, error, and Taylor plots (employing various metric values) is essential for understanding the estimation accuracy of developed models and comparing their performance to introduce the optimal one for real-world applications. Figure 9 indicates scattered representations of the comparison between predicted values by each pair of models in single, hybrid (optimized with ROA and DMO), and ensemble groups. The reported data points are according to RMSE (dispersion controller) and R2 (accordance with the center line of ideal prediction) values. Moreover, two lines are drawn below and above the centerline for 10% and 20% over and underestimation. Considering single models, RFR with higher R2 and lower RMSE values performed better than DTR, but both of the models had higher than 20% misestimations, which proved the necessity of optimization of traditional regression models. Turning to the comparison of hybrid models, it is evident that ROA was more powerful than DMO in optimizing both models (especially RFR), being more saucerful in restricting prediction datapoints between thresholds of -20% and 20%. ROA reduced the error value of DTR and RFR by 1.3 and 1.6 million dollars (10% of average values presented for players of each studied league in Table 1). Optimizing models with an ensemble form of two optimizers predicted market values of players with high R2 and accuracy, which is the most reliable model for predictions by any dataset with various ranges and order.

Managerial decisions to buy players with high salaries in the next season based on the performance of the players in the current season are riskier than decision-making about less expensive players. Hence, data points near the position to the center line, especially in the case of predictions by RFRO, DTRO, RFRD, and DTRD, indicate they are capable tools for reliable decision-making in the competitive transfer market of footballers.

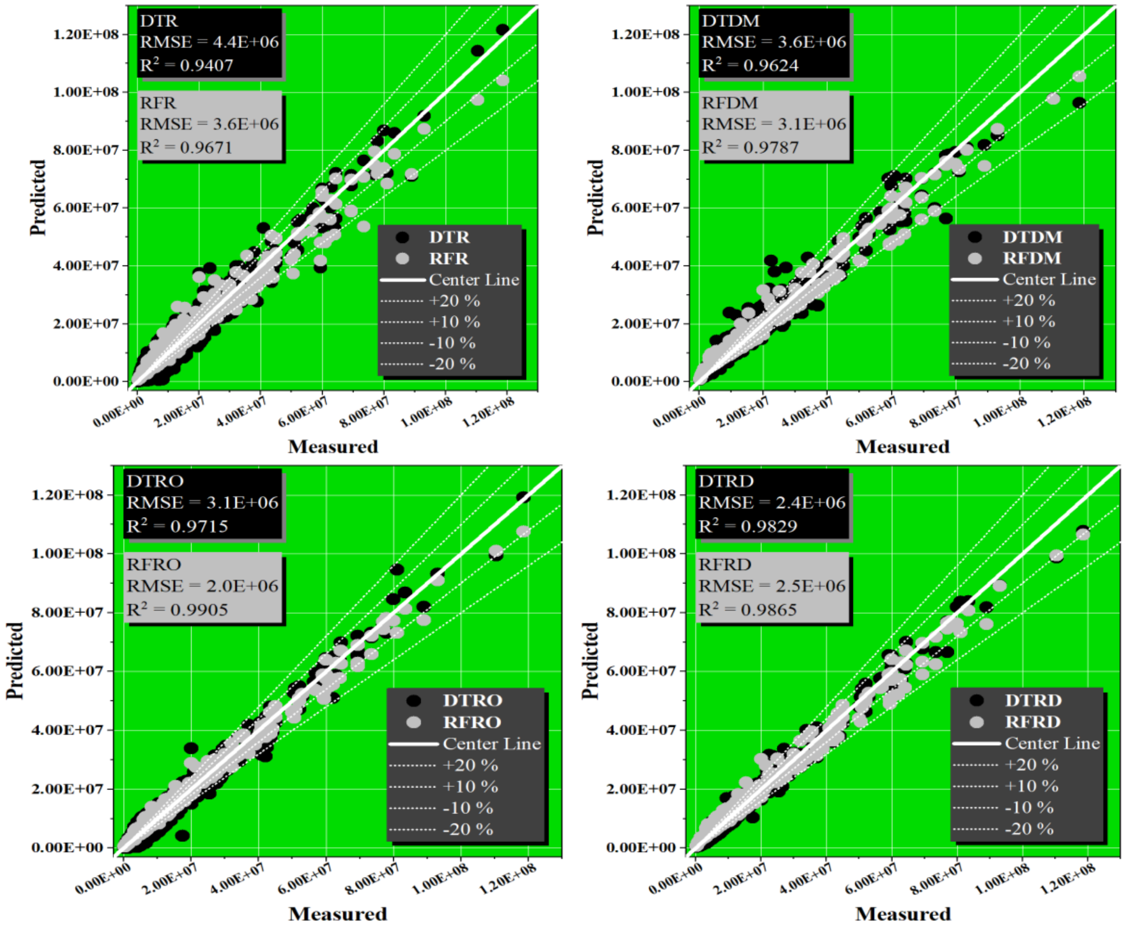

Figure 9. Scatter plot for comparison between measured market values and predicted values.

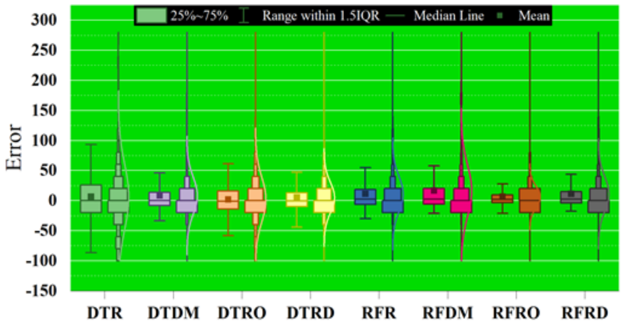

Figure 10. Comparative prediction error for single, hybrid, and ensemble models.

The sequential pattern in the box plot representation of prediction errors in Figure 10 compares prediction errors of developed models to further illustrate the development in the accuracy of estimations by employing the most recent hybrid and ensemble methods compared to single traditional models. Low error ranges of all hybrid models were visible compared with their single counterparts, especially in the case of models optimized with ROA and an ensemble of two optimizers. RFRO had the nearest mean error value to zero percent and the narrowest range for 25% to 75% of error values. RFRD and DTRD came in the second and third positions of the ranking. It is worthwhile to note that the mean error indicator (square) for all models was top of the median line, which represents an overestimation of prediction models that may create a margin of safety for club managers with a low probability of lack of funds in next season.

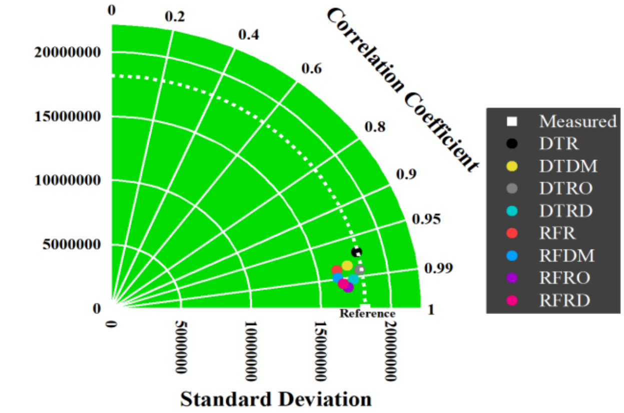

Figure 11. comparison between estimation accuracy of developed models.

In the final part of this section, the Taylor plot in Figure 11 illustrates a comparison of the prediction performance of developed models considering Correlation Coefficient, Standard deviation, and RMSE values. The measured value is the benchmark and position of each model's performance indicator (colored circles) in comparison with the real value is the determining factor of prediction accuracy. The standard deviation of predicted values by DTR and DTRO was less than other predictors, which led to their alignment with the reference line, but their accuracy of predictions based on R2 and RMSE values was low (distance from real market values). RFRO and RFRD were the most precise estimators with a correlation coefficient of higher than 99%.

5.2. Evaluation of Players' Value in the Transfer Market of Various Leagues

To evaluate the generalization performance of estimation models, their performance is examined through error values representation for five different European leagues. Figure 12 shows the prediction error of models for various leagues' transfer markets. RFRO and DTR had the smallest and largest range of error values for all players regardless of their playing leagues. Of course, the low error range in the case of the Premier League with the highest average and minimum market values (Table 1) is detectable in all RFR-based models, indicating precise estimations for clubs with valuable players. All these comparative results revealed that managers of well-known clubs (in first-rank leagues) could rely on predictions of these models (especially the most optimal models of RFRO and RFRD) with an accuracy of about 90% in choosing players with optimal values for the coming season based on their performance to arrange best teams regarding allocated funds.

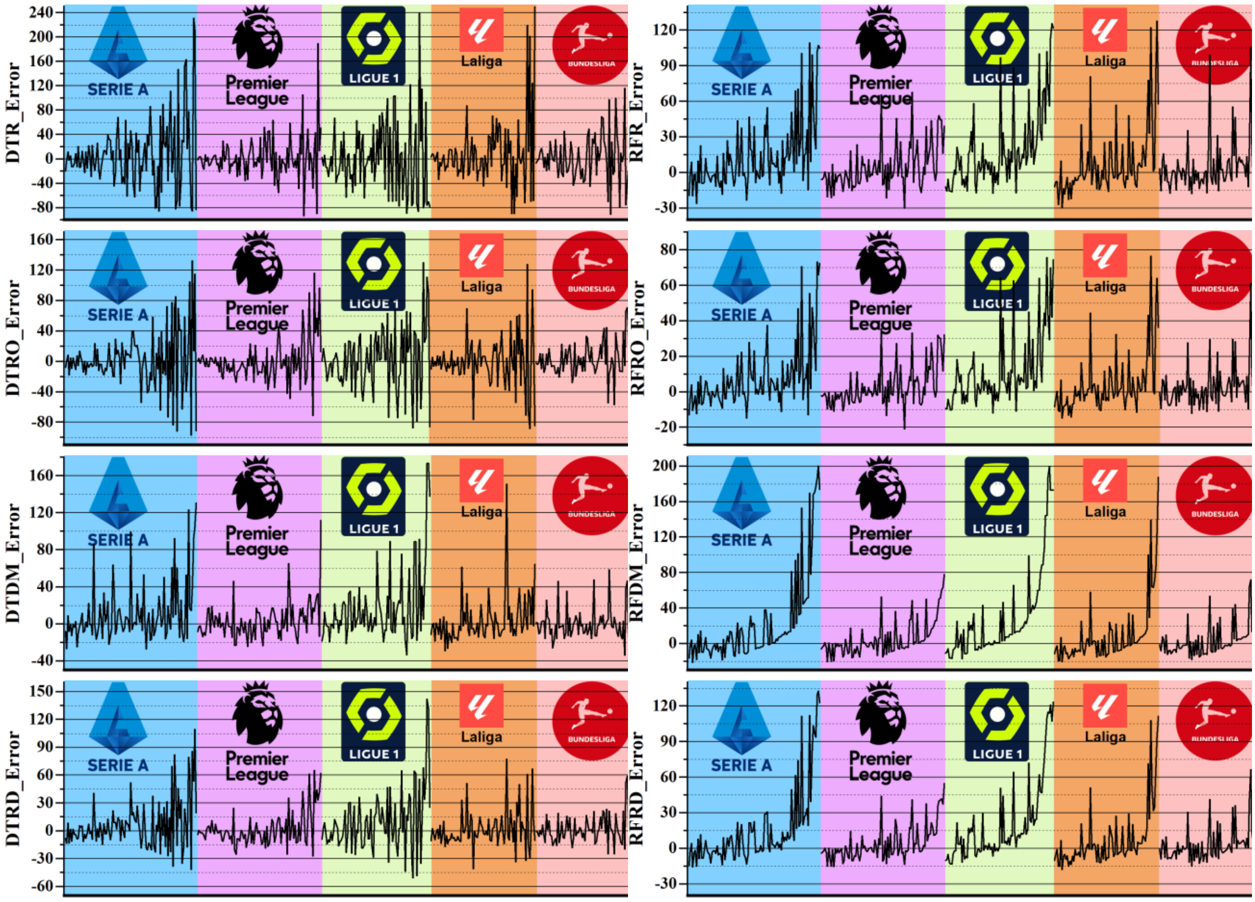

Figure 12. Prediction error of models for various leagues' transfer market.

5.3. Comparison with Prediction Models in Existing Literature

Several studies in the existing literature have investigated the market valuation of football players. This section undertakes a comparative analysis, focusing on the R2 metric, between the top-performing model developed in this study and those documented in prior research. Although all developed models in this study presented higher R2 values than presented models existing in the literature, based on reports in Table 6, the R2 results related to RFRO were 5%-25% better than the prediction results of best-developed models in previous literature.

Table 6. Comparative analysis of developed model's prediction accuracy with models in previous studies.

Reference	Best prediction model	R2
[9]	SVR-PSO	0.74
[34]	XGB	0.77
[5]	RFR	0.95
Present work	RFRO	0.9905
	DTRO	0.9865
	RFRD	0.9715
	DTRD	0.9828

5.4. Contributions of the Study and Model Integration in Football

The contribution of this study lies in its comprehensive evaluation of predictive models for estimating football player market values, addressing both performance metrics and practical implications. The analysis of the obtained results highlights the superiority of RFR-based models over DTR-based counterparts in explaining variance. Notably, the exceptionally high R2 value of approximately 98% for ensemble models underscores the efficacy of combining multiple models for improved accuracy. Moreover, these prediction models can offer valuable guidance for decision-makers in the football industry, enabling optimized player selection and budget allocation strategies based on historical performance data.

Integrating the proposed models into practical decision-making workflows for football clubs may encounter several challenges, along with strategies to address them effectively. Firstly, one challenge could be the availability and quality of data. While the study utilizes data from FIFA 19 video games and real-world statistics, ensuring access to up-to-date and comprehensive player data from various leagues worldwide might pose difficulties. To mitigate this, establishing partnerships with data providers and leveraging advanced data scraping techniques could help in gathering relevant data efficiently. Secondly, the complexity of the models themselves may present a challenge in terms of interpretation and implementation. Football clubs may lack the expertise or resources to understand and apply sophisticated ML algorithms like RFR. Providing user-friendly interfaces and documentation, along with offering training and support, can aid clubs in effectively utilizing these models. Additionally, ensuring transparency in model outputs and decision-making processes is crucial for gaining trust and buy-in from club stakeholders.

5.5. Future Works

The dataset in this study was related to top-ranked leagues of Europe, so writers suggest collecting a more diverse dataset related to lower-ranked or football leagues of other continents; moreover, in all discussions in this study playing position of players was ignored as an effective factor on their market values representing the gap in studies. Of course, in the case of prediction models' development, other types of regression models (rather than tree-based models utilized in this study) and novel approaches of two or more models ensembling will give valuable insights into the application of ML predictions in the case of football players transfer market.

6. Conclusions

This study explored the predicting football player market values across prominent European Leagues (Serie A, Premier League, League 1, La Liga, and Bundesliga), employing a comprehensive dataset that integrates real-world statistical datasets from FIFA 19 and Real-World Statistics. The research underscores the dynamic nature of player valuation, influenced by multifaceted factors such as skill proficiency, age, on-field performance, contractual status, and market demand in leagues, which may face the dataset with multicollinearity. Data engineering, such as cleaning outliers, eliminating unnecessary factors, and extracting pertinent information to create a robust dataset to generate a more comprehensive and relevant dataset. A critical component of this process was feature selection, where the VIF approach was employed to address multicollinearity issues within the dataset, identifying variables that exhibit substantial correlation with others. By reducing the number of imperative features based on their correlation, VIF ensured a more streamlined and efficient dataset of 459 samples and 33 features, facilitating the subsequent prediction models. Prediction study conducted by advanced hybrid ML techniques, specifically DTR and RFR models optimized with the ROA and DMO. The evaluation across major European leagues reveals the superiority of RFR-based models, with RFRO standing out as a high-performing predictor (2 million dollars misestimation, which is just 12% of average market values). The ensemble models RFRD and DTRD further demonstrate their efficacy with reliable prediction capabilities. The presented models had precise predictions for all leagues, which showed the generality of such models, which enables club managers to have an accurate estimate of the next season's budget for the purchase of players. Also, enabling coaches to choose the best combination of players that fits the allocated budget in addition to the necessary abilities.

References

[1] Agushaka J., Ezugwu A., and Abualigah L., "Dwarf Mongoose Optimization Algorithm," *Computer Methods in Applied Mechanics and Engineering*, vol. 391, pp. 114570, 2022. https://doi.org/10.1016/j.cma.2022.114570

[2] Ahmad A., Farooq F., Niewiadomski P., Ostrowski K., Akbar A., Aslam F., and Alyousef R., "Prediction of Compressive Strength of Fly Ash Based Concrete Using Individual and Ensemble Algorithm," *Materials*, vol. 14, no. 4, pp. 1-21, 2021. https://doi.org/10.3390/ma14040794

[3] Akinola O., Ezugwu A., Oyelade O., and Agushaka J., "A Hybrid Binary Dwarf Mongoose Optimization Algorithm with Simulated

Annealing for Feature Selection on High Dimensional Multi-Class Datasets," *Scientific Reports*, vol. 12, no. 1, pp. 1-22, 2022. https://www.nature.com/articles/s41598-022-18993-0

[4] Akinwande M., Dikko H., and Samson A., "Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable (s) in Regression Analysis," *Open Journal of Statistics*, vol. 5, no. 7, pp. 754-767, 2015. DOI:10.4236/ojs.2015.57075

[5] Al-Asadi M. and Tasdemır S., "Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques," *IEEE Access*, vol. 10, pp. 22631-22645, 2022. DOI:10.1109/ACCESS.2022.3154767

[6] Angel B. and Gasparetto T., *Routledge Handbook of Football Business and Management*, Routledge, 2018. https://doi.org/10.4324/9781351262804

[7] Baldi R., *They Always Score: The Unforgettable, Improbable, Iconic Story of Manchester United's Treble Winners*, Birlinn Ltd, 2023. https://www.amazon.co.uk/They-Always-Score-Unforgettable-Improbable/dp/191353895

[8] Basirat M., Khajeheian D., and Arbatani T., "A Theoretical Model for Identifying Media Value of Football Players in Iranian Professional League," *Sport Management Studies*, vol. 11, no. 57, pp. 121-140, 2019. https://doi.org/10.22089/smrj.2019.7317.2550

[9] Behravan I. and Razavi S., "A Novel Machine Learning Method for Estimating Football Players' Value in the Transfer Market," *Soft Computing*, vol. 25, no. 3, pp. 2499-2511, 2021. https://doi.org/10.1007/s00500-020-05319-3

[10] Breiman L., "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001. https://doi.org/10.1023/A:1010933404324

[11] Breiman L., Friedman J., Olshen R., and Stone C., *Classification and Regression Trees*, CRC Press, 1984. https://www.academia.edu/5867603/Classification_and_Regression_Trees

[12] Carmichael F., Rossi G., and Thomas D., "Production, Efficiency, and Corruption in Italian Serie A Football," *Journal of Sports Economics*, vol. 18, no. 1, pp. 34-57, 2017. https://doi.org/10.1177/1527002514551802

[13] Coates D. and Parshakov P., "The Wisdom of Crowds and Transfer Market Values," *European Journal of Operational Research*, vol. 301, no. 2, pp. 523-534, 2022. https://doi.org/10.1016/j.ejor.2021.10.046

[14] Elliott R., *The English Premier League: A Socio-Cultural Analysis*, Taylor and Francis, 2017. https://doi.org/10.4324/9781315636696

[15] Erdal H., "Two-Level and Hybrid Ensembles of Decision Trees for High Performance Concrete Compressive Strength Prediction," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 7, pp. 1689-1697, 2013. https://doi.org/10.1016/j.engappai.2013.03.014

[16] Felipe J., Fernandez-Luna A., Burillo P., De la Riva L., Sanchez-Sanchez J., and Garcia-Unanue J., "Money Talks: Team Variables and Player Positions that most Influence the Market Value of Professional male Footballers in Europe," *Sustainability*, vol. 12, no. 9, pp. 1-8, 2020. https://doi.org/10.3390/su12093709

[17] Fieldsend D., *The European Game: The Secrets of European Football Success*, Arena Sport, 2017. https://www.casematepm.com/9781909715486/the-european-game/

[18] Franceschi M., Brocard J., Follert F., and Gouguet J., "Football Players in Light of Economic Value Theory: Critical Review and Conceptualisation," *Managerial and Decision Economics*, vol. 45, no. 2, pp. 896-920, 2023. https://doi.org/10.1002/mde.4039

[19] González-Rodenas J., Moreno-Pérez V., López-Del Campo R., Resta R., and Del Coso J., "Evolution of Tactics in Professional Soccer: An Analysis of Team Formations from 2012 to 2021 in the Spanish LaLiga," *Journal Human Kinetics*, vol. 87, pp. 207-216, 2023. DOI:10.5114/jhk/167468

[20] Goswami S. and Chakrabarti A., "Feature Selection: A Practitioner View," *International Journal of Information Technology and Computer Science*, vol. 6, no. 11, pp. 66-77, 2014. DOI:10.5815/ijitcs.2014.11.10

[21] Hamil S., "The German Football Bundesliga," *Birkbeck College*, pp. 1-17, 2014. https://www.academia.edu/10298953/THE_GERMAN_FOOTBALL_BUNDESLIGA

[22] He M., Cachucho R., and Knobbe A., "Football Player's Performance and Market Value," *in Proceedings of the Machine Learning and Data Mining for Sports Analytics ECML/PKDD Workshop*, Porto, pp. 87-95, 2015. https://api.semanticscholar.org/CorpusID:39624891

[23] Herm S., Callsen-Bracker H., and Kreis H., "When the Crowd Evaluates Soccer Players' Market Values: Accuracy and Evaluation Attributes of an Online Community," *Sport Management Review*, vol. 17, no. 4, pp. 484-492, 2014. https://doi.org/10.1016/j.smr.2013.12.006

[24] Horn C., An Exploratory Study into Select Technical Key Performance Indicators and Estimated Transfer Fees in the 2nd Division of the German Bundesliga, Master's Thesis, Stellenbosch University, 2023. https://scholar.sun.ac.za/server/api/core/bitstreams/c52a0036-07f6-45e0-8a24-0c26bda7a3b2/content

[25] Isikdemir E., Ozkurkcu S., and Ozer S.,

"Technical Analysis of Goals Scored in 3 Different European Leagues in the 2020-2021 Football Season," *Journal of Sport Sciences Researches*, vol. 8, no. 3, pp. 458-472, 2023.

[26] Juuri S., Predicting the Results of NFL Games Using Machine Learning, Master's Thesis, Aalto University, 2023. https://aaltodoc.aalto.fi/server/api/core/bitstreams/80b6e0d0-f5d1-4c19-abd3-667ee40d9c93/content

[27] Karbassi A., Mohebi B., Rezaee S., and Lestuzzi P., "Damage Prediction for Regular Reinforced Concrete Buildings Using the Decision Tree Algorithm," *Computers and Structures*, vol. 130, pp. 46-56, 2014. https://doi.org/10.1016/j.compstruc.2013.10.006

[28] Lepschy H., Wasche H., and Woll A., "Success Factors in Football: An Analysis of the German Bundesliga," *International Journal of Performance Analysis in Sport*, vol. 20, no. 2, pp. 150-164, 2020. https://www.tandfonline.com/doi/full/10.1080/24748668.2020.1726157

[29] Liaw A. and Wiener M., "Classification and Regression by RandomForest," *R News*, vol. 2, no. 3, pp. 18-22, 2002. https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf

[30] Lowe S., *Fear and loathing in La Liga: Barcelona, Real Madrid, and the World's Greatest Sports Rivalry*, Bold Type Books, 2014. https://www.amazon.com/Fear-Loathing-Liga-Barcelona-Greatest/dp/1568584504

[31] Mahareek E., Cifci M., El-Zohni H., and Desuky A., "Rhizostoma Optimization Algorithm and its Application in Different Real-World Optimization Problems," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 4, pp. 4317-4338, 2023. DOI:10.11591/ijece.v13i4.pp4317-4338

[32] Majewski S., "Identification of Factors Determining Market Value of the most Valuable Football Players," *Central European Management Journal*, vol. 24, no. 3, pp. 91-104, 2016. DOI:10.7206/jmba.ce.2450-7814.177

[33] Matheson V., European Football: A Survey of the Literature, Williams College, Department of Economics Williamstown, 2003.

[34] McHale I. and Holmes B., "Estimating Transfer Fees of Professional Footballers Using Advanced Performance Metrics and Machine Learning," *European Journal of Operational Research*, vol. 306, no. 1, pp. 389-399, 2023, https://doi.org/10.1016/j.ejor.2022.06.033

[35] Metelski A., "Factors Affecting the Value of Football Players in the Transfer Market," *Journal of Physical Education and Sport*, vol. 21, no. 2, pp. 1150-1155, 2021. https://efsupit.ro/images/stories/aprilie2021/Art%20145.pdf

[36] Peters J., De Baets B., Verhoest N., Samson R., Degroeve S., De Becker P., and Huybrechts W., "Random Forests as a Tool for Ecohydrological Distribution Modelling," *Ecological Modelling*, vol. 207, no. 2-4, pp. 304-318, 2007. https://doi.org/10.1016/j.ecolmodel.2007.05.011

[37] Podzemsky L., Analysis of Investments and Market Value of Football Clubs, Bachelor's Thesis, Charles University 2022. https://dspace.cuni.cz/bitstream/handle/20.500.11956/175470/130344478.pdf?sequence=1&isAllowed=y

[38] Prinz A. and Thiem S., "Value-Maximizing Football Clubs," *Scottish Journal Political Economy*, vol. 68, no. 5, pp. 605-622, 2021. https://doi.org/10.1111/sjpe.12282

[39] Putra M., Dewi D., Putri W., Hendrowati R., and Kurniawan T., "Contributed Factors in Predicting Market Values of Loaned out Players of English Premier League Clubs," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 9, pp. 359-365, 2023. DOI:10.14569/IJACSA.2023.0140939

[40] Rodriguez-Galiano V., Sanchez-Castillo M., Chica-Olmo M., and Chica-Rivas M., "Machine Learning Predictive Models for Mineral Prospectivity: An Evaluation of Neural Networks, Random Forest, Regression Trees and Support Vector Machines," *Ore Geology Reviews*, vol. 71, pp. 804-818, 2015. https://doi.org/10.1016/j.oregeorev.2015.01.001

[41] Rossi G., Tanna G., and Addesa F., "Production, Efficiency and Corruption in Italian Serie A: A DEA Analysis," Birkbeck, University of London, vol. 9, no. 1, pp. 1-24, 2016. https://eprints.bbk.ac.uk/id/eprint/18395/

[42] Sadoun A., Najjar I., Alsoruji G., Wagih A., and Abd Elaziz M., "Utilizing a Long Short-Term Memory Algorithm Modified by Dwarf Mongoose Optimization to Predict Thermal Expansion of Cu-Al2O3 Nanocomposites," *Mathematics*, vol. 10, no. 7, pp. 1-17, 2022. https://www.mdpi.com/2227-7390/10/7/1050

[43] Sener I. and Karapolatgil A., "Rules of the Game: Strategy in Football Industry," *Procedia-Social and Behavioral Sciences*, vol. 207, pp. 10-19, 2015. https://doi.org/10.1016/j.sbspro.2015.10.143

[44] Sengupta S., "Understanding La Liga: Are Match Performances and Player Market Value Related?," *International Research Journal of Nature Science and Technology*, vol. 2, no. 6, pp. 1-11, 2020. https://scienceresearchjournals.org/IRJNST/2020/volume-2%20issue-6/irjnst-v2i6p101.pdf

[45] Singh G. and Panda R., "Daily Sediment Yield Modeling with Artificial Neural Network Using 10-fold Cross Validation Method: A Small Agricultural Watershed, Kapgari, India," *International Journal of Earth Sciences and*

Engineering, vol. 4, no. 6, pp. 443-450, 2011. file:///C:/Users/user/Downloads/Daily_Sediment_Yield_Modeling_with_Artificial_Neur.pdf

[46] Swanepoel M. and Swanepoel J., "The Correlation between Player Valuation and the Bargaining Position of Clubs in the English Premier League (EPL)," *International Journal of Economics and Finance Studies*, vol. 8, no. 1, pp. 209-225, 2016. https://www.sobiad.org/eJOURNALS/journal_IJEF/archieves/IJEFS2016_1/Paper74_Swanepoel_Swanepoel.pdf

[47] Tomlinson A. and Young C., *German Football: History, Culture, Society*, Routledge, 2006. http://ndl.ethernet.edu.et/bitstream/123456789/24867/1/Alan_Tomlinson_2006.pdf

[48] Wagner F., Preuss H., and Könecke T., "A Central Element of Europe's Football Ecosystem: Competitive Intensity in the 'Big Five," *Sustainability*, vol. 13, no. 6, pp. 3097, 2021. https://doi.org/10.3390/su13063097

[49] Wang Y., Tarakci H., and Prybutok V., "Model Comparison of Regression, Neural Networks, and XGBoost as Applied to the English Premier League Transfer Market," *International Journal of Sport Management and Marketing*, vol. 23, no. 6, pp. 543-559, 2023. https://doi.org/10.1504/IJSMM.2023.133786

[50] Zaib R. and Ourabah O., "Large Scale Data Using K-Means," *Mesopotamian Journal of Big Data*, vol. 2023, pp. 36-45, 2023. https://doi.org/10.58496/MJBD/2023/006

Yu Sun, Chinese Han nationality, born in Siping City, Jilin Province, China, lecturer of Jilin Agricultural University, graduated from Northeast Normal University, master degree, mainly engaged in sports training, physical education teaching research.

Kepeng Gu, Chinese Han nationality, born in Liaoyuan City, Jilin Province, China, associate professor of Jilin Agricultural University, graduated from Northeast Normal University, master degree, mainly engaged in sports training, physical education teaching research.