Football Player Value Prediction: Comparing Machine Learning Models with Cross-Validation

Yitong Kong*

Department of Mathematics, Sorbonne University, Paris, 75005, France

*Corresponding author: Yitong.kong@etu.sorbonne-universite.fr

Abstract. The project mainly focuses on the evaluation of different models used in order to predict the current value of football players based on a dataset from Kaggle. Three models-random forest, gradient boosting, and ridge regression-are being examined using key indicators such as R-squared (R2), root mean square error (RMSE), and mean absolute percentage error (MAPE). A cross-validation process is applied to ensure the robustness of the model evaluations. Among these models, the gradient boosting model is the most suitable since it provides the lowest RMSE and the highest R2 indicating high accuracy. A young, robust, and healthy player who has had a high market value in the past or present is more likely to have a high current value. The research aims to provide team managers with a relatively accurate model to predict player's market value. At the same time, this study can help players understand which factors impact the most in terms of current value, encouraging them to improve themselves in certain areas.

Keywords: Current value; cross-validation; gradient boosting; random forest; ridge regression.

1. Introduction

Every year, the investment in the world of football is massive. Unlike 20 years ago, nowadays the assets that the top football club owns are already an astronomical figure. Likewise, the market value of certain players has almost tripled. For instance, during the summer transfer window of 2024, Arsenal spent around 100 million pounds, and the most expensive player Kylian Mbappé values 180 million euros today.

Due to this trend, it becomes more and more important for football managers to evaluate the current value of a certain player they are interested in so that they can offer a reasonable contract when competing against others. In this case, different methods and models of machine learning can be trained and eventually applied to predict the current value of a certain player precisely. The goal of this paper is to figure out which model is the most appropriate for the prediction of the current value of football players.

The dataset is collected from a shared file on Kaggle, which is undertaken to create prediction models for the transfer fees of football players. The dataset gathers 20 top-level leagues and the data from 10,755 players in 2 seasons 2021-2022 and 2022-2023. The dataset provides a list of characteristics of football players for instance goals, assists, etc. Since the estimation of transfer fee involves even more aspects of a player, for example, the remaining years of the current contract, or even the relationship between the player and the club, the prediction of the current value of a player, which mainly depends on the performance of the player on the pitch, will be more precise to some extent. That is the reason why a current value prediction will be preferable in the project.

In recent years, using different models to predict the current value of athletes has become a trendy popularity. Several studies have already adopted different approaches to predicting transfer fees or market value using various datasets and models.

Firstly, the work of Poli et al. established an econometric model that analyses the key factors which impact largely on the transfer fee of footballers [1]. multiple linear regression (MLR) is adopted. The model is trained based on over 2,000 player transfers across the major leagues in Europe during a decade, from 2012 to 2021. The applied data is well-constructed and verified since cross-referencing from reliable sources for instance Transfermarkt is utilized.

One vital factor in determining the transfer fees, which nevertheless is often neglected by previous research, is the current contract according to Poli et al. [1]. Their findings suggest that the estimation of clubs on a certain player not only depends on the player's in-pitch performance but also the details of the contract, for example, the remaining duration of the contract, which often serves as a critical weapon during negotiations.

As indicated by Poli et al., the gap between transfer value and the current value of a player is relatively large, as the transfer fee, which is certainly related to the performance of the player, sometimes reflects the financial or strategic decision made by the club [1]. The goal of the work of Poli et al. is to help clubs make better decisions regarding the purchase or the negotiation of players by using predictive models [1].

Due to missing data and the difficulty of quantifying some of the data for example the remaining duration of the contract, this paper chose to analyze current value rather than transfer fees.

Secondly, the work of Anjun et al. provides a new approach to predicting the market value of football players, diverging from traditional econometric models [2]. Anjun et al. chose to draw the data from a popular database Sofifa, which is a website that extracts data from FIFA games. That is, its data comes from FIFA games, which probably gives rise to some errors or distortions from the real world since the data itself is sometimes predicted by models that FIFA games used. Their research employs four different regression techniques-MLR, random forest, decision trees, and linear regression-to identify the most accurate predictive model.

It is shown that the random forest model is superior to other models thanks to its high accuracy and low error: the optimized random forest achieves an RMSE of nearly 1.4 times better than other regression models.

Thirdly, the work of He used primarily linear models to predict the current value of footballers. The database is extracted from Transfermarkt and Wikipedia [3]. The former is considered a relatively reliable source since it enjoys a great reputation worldwide while the latter is less convincing for the fact that everyone can edit. The novelty of He's work is that the 16 independent variables that may affect the market value are divided into three major categories: personal information (e.g., age and position), performance metrics (e.g., goals and assists), and calculated ratios (e.g. goals per game) [3]. This classification facilitates the analysis of reflecting the different aspects influencing a player's market value. The work of He utilizes various kinds of linear models including ordinary least squares (OLS), ridge regression, and principal component regression (PCR) [3].

One of the findings of He's work is that not only the performance metrics but also the background of the player like the position and the reputation of the club the player plays for have a profound influence on the market value of the player [3]. However, since most of the models are driven by linear techniques, some of the non-linear metrics may not be correctly estimated in the predicting of the final current value of players.

On a different note, the work of McHale et al. introduces a top-end machine learning algorithm, the XGBoost to model transfer fees [4]. The dataset adopted in this work is partly from Sofifa, a crowd-sourced platform where everyone can rate or edit a player, and partly from advanced tracking of football matches, which provides unique advanced data e.g. expected goals and expected assists that evaluate the opportunity seizing ability of a player. The prediction of XGBoost is rather accurate, with an R2 score reported as 0,77 and a Mean Absolute Error (MAE) of 3.60.

The study of McHale et al. is also able to identify transfers that are valuable compared to others [4]. For example, some clubs like Liverpool and Atletico Madrid excel at making good deals or bargaining since they can always offer a price that is lower than the estimated transfer fee of players.

Lastly, in the work of Lee et al., an optimized LightGBM model is applied to the prediction [5]. The data is obtained from Sofifa and WhoScored which is a website providing advanced football statistics. According to Lee et al., the LightGBM outpaces the performance of traditional models like random forest since a significantly lower RMSE of LightGBM is reported compared to that of random forest [5]. This reduction in error is achieved through hyperparameter optimization using the tree-structured parzen estimator (TPE), which further fine-tunes the model for higher accuracy.

After reading these sampled theses, a cross-validation process seems likely to be forgotten. However, in this project, 3-fold cross-validation is adapted to ensure the robustness of models. Furthermore, a reliable dataset from Kaggle is being used in this paper, which ensures accuracy. Compared to some of the sampled papers, this paper not only introduces linear regression models but also includes non-linear models which are more precise concerning the depiction of some potential relationships.

2. Methodology

2.1. Data Source

The dataset adopted in this project is a public project shared on Kaggle called "Football players' transfer fee prediction dataset" whose usability is rated at 10.0/10.0, showing its completeness and credibility.

Table 1. Variables of the adopted dataset

Variable	Type	Value Range
Player	Object	-
Team	Object	-
Name	Object	-
Position	Object	-
Height	Int64	[156; 206]
Age	Int64	[15; 43]
Appearance	Int64	[0; 107]
Goals per game	Float64	[0.0; 5.0]
Assists per game	Float64	[0.0; 4.0]
Yellow cards per game	Float64	[0.0; 1.93]
Second yellow cards per game	Float64	[0.0; 1.0]
Red cards per game	Float64	[0.0; 0.96]
Goals conceded per game	Float64	[0.0; 9.0]
Clean sheet	Float64	[0.0; 1.0]
Minutes played	Int64	[0; 9510]
Days injured	Int64	[0; 1570]
Games injured	Int64	[0; 339]
Award	Int64	[0; 92]
Current value	Int64	[10,000; 180,000,000]
Highest value	Int64	[10,000; 200,000,000]
Position encoded	Int64	{1; 2; 3; 4}
Winger	Boolean	{0; 1}

According to Table 1, the dataset includes basic football players' basic information including age, height, position, club, and awards obtained. The performance metrics of each player (e.g. appearance, goals, assists, yellow cards, second yellow cards, red cards, goals conceded, clean sheets, and injuries) during 2 seasons (season 2021-2022 and 2022-2023) are also included. The factors mentioned above are considered variables in machine learning, and the ultimate goal of this project, or namely the variable this project wants to predict and compare with real statistics, is the current value of footballers. The dataset contains 10,755 players in 20 top-tier leagues worldwide. However, by checking the dataset manually, some players' current values are marked as 0, which is impossible in real life, therefore, these data are removed to prevent significant prediction inaccuracies.

2.2. Method Introduction

In this study, the models chosen to be utilized during machine learning are random forest, gradient boosting, and ridge regression. First of all, since the random forest is a model that is non-parametric and flexible, according to Horning, this model can deal with not only continuous but also categorical variables, which are extremely important in the prediction of the market value of football players based on their complex backgrounds, for instance, their position or club [6]. As for ridge regression, it helps manage large numbers of predictors more efficiently than linear regression according to the article by Firinguetti et al. [7]. Last, thanks to its sequential learning feature, gradient boosting builds models sequentially, namely it generates new models after correcting the errors of the previous ones, which will give rise to an even more accurate prediction than random forest [8]. Thanks to this feature, gradient boosting is capable of dealing with complex datasets, outperforming normal linear regression models.

2.3. Data processing

Firstly, in the provided dataset, some elements are irrelevant to predicting the current value of athletes such as names, which is why these columns are abandoned. Additionally, the rows where the current value is 0 are all removed because it is not realistic. Subsequently, a log transformation to the current value is applied to reduce skewness because a few players have extremely high values, whereas most players have relatively low values, so later on the prediction will be probably distorted. Then, some variables (e.g. team and position) are transformed into dummy variables for quantification and a better interpretation of models. Afterward, the variables are standardized and split into training and testing sets (80/20 split). In the following step, three models, random forest, gradient boosting, and ridge regression (with alpha=10), are trained along with a 3-fold Cross-validation which ensures the robustness of the prediction as well as the efficiency of the code. The ridge regression with a parameter (alpha equal to 10) prevents the model from overfitting while sacrificing some accuracy.

In the interest of evaluating the accuracy of each model, 3 indicators are being utilized: RMSE, R2, and MAPE. Compared to MAE, RMSE can highlight differences in model performance more sharply because it emphasizes larger errors, which makes it easier to evaluate models when dealing with extreme cases, such as players with very high actual values in this project [9]. On the other hand, MAPE expresses the error as a percentage of actual values bringing an obvious way to evaluate the accuracy of models. The range of R2 is from 0 to 1, the closer to 1, the more precise the prediction delivers [10].

Additionally, the obtained data are also visualized, with comparisons made between each model's R^2, RMSE, and MAPE, as well as scatter plots comparing the predicted values to the actual data for each model to render better legibility.

Finally, since random forest and gradient boosting are both tree-based models that can deliver relatively accurate predictions, their top five most important features to determine the models are also plotted.

3. Results and Discussion

3.1. Error Analysis

After processing the dataset and running the code, the result is obtained as below.

Table 2. Performance metrics comparison of prediction models

Models	RMSE	R^2	MAPE (%)
Random forest	0.463	0.924	2.222
Gradient boosting	0.450	0.928	2.294
Ridge regression	0.911	0.705	5.321

As shown in Table 2, it can be seen that among these three models, gradient boosting represents the lowest RMSE and the highest R2 indicators, while random forest stands for the lowest MAPE.

3.2. Visualization

It can be seen that among these three models, gradient boosting represents the lowest RMSE and the highest R2 indicators, while random forest stands for the lowest MAPE.

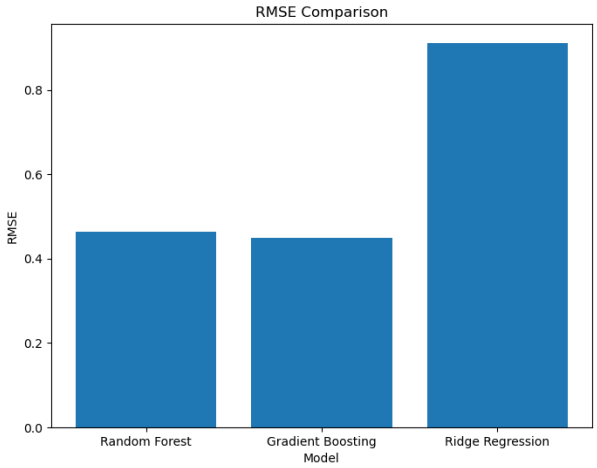

Fig. 1 RMSE comparison

In terms of RMSE, it is clearly shown in Figure 1 that gradient boosting represents the lowest RMSE, followed closely by random forest. Whereas ridge regression possesses a higher RMSE showing that its prediction deviates from the actual value.

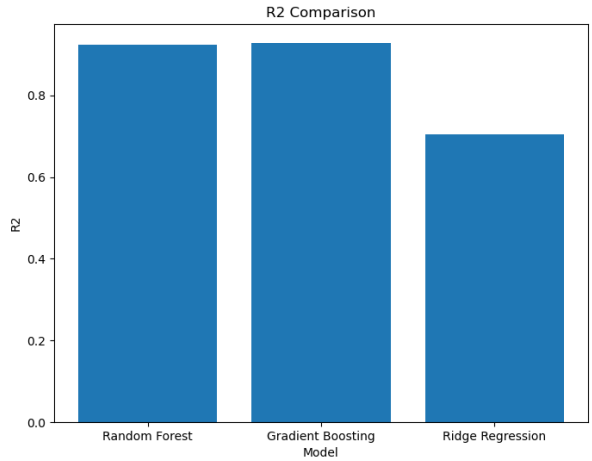

Fig. 2 R^2 comparison

According to Figure 2, the R^2 indicator reflects similar results following RMSE. The value of random forest and gradient boosting is closer to 1 competing with that of ridge regression, ensuring a high accuracy of the first two.

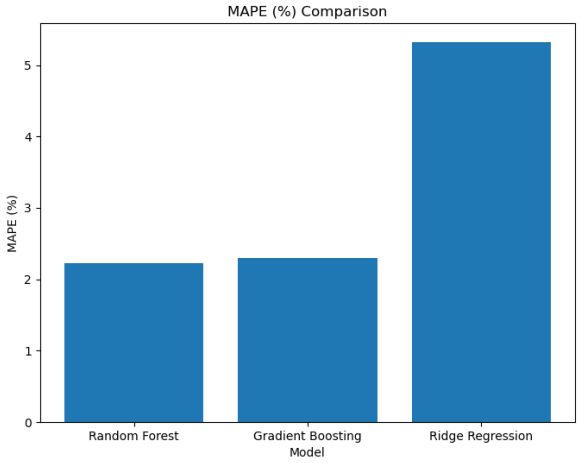

Fig. 3 MAPE (%) comparison

Thanks to Figure 3, it is indicated that random forest and gradient boosting both achieve a MAPE of around 2%. In contrast, ridge regression has a significantly high MAPE (5% approximately), proving a poorer performance.

Additionally, to more clearly demonstrate the accuracy of each model, the predicted current values were compared with the actual data, and scatter plots were created.

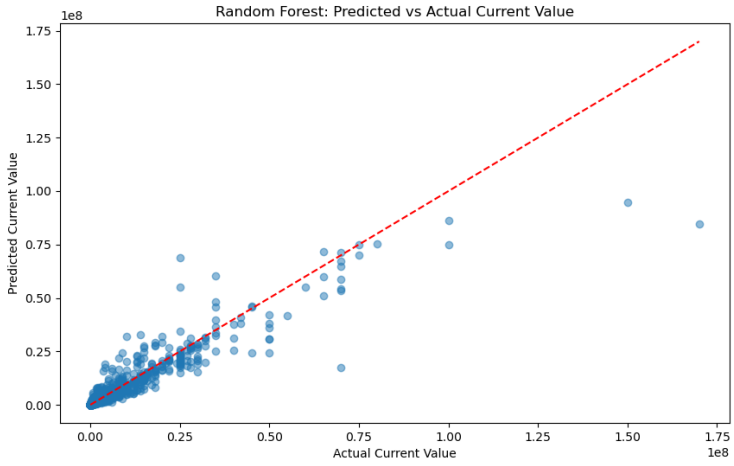

Fig. 4 Random forest: predicted value compared with the actual value

When comparing the actual value and the predicted value through random forest, which is visualized in Figure 4, it can be seen that despite some slight deviations, the model fits well concerning low-value players, which compose the majority of the dataset. Nonetheless, due to the scarcity of high-value players (only a few elite footballers merit and top-tier clubs can afford them), the model always tends to underestimate their market value.

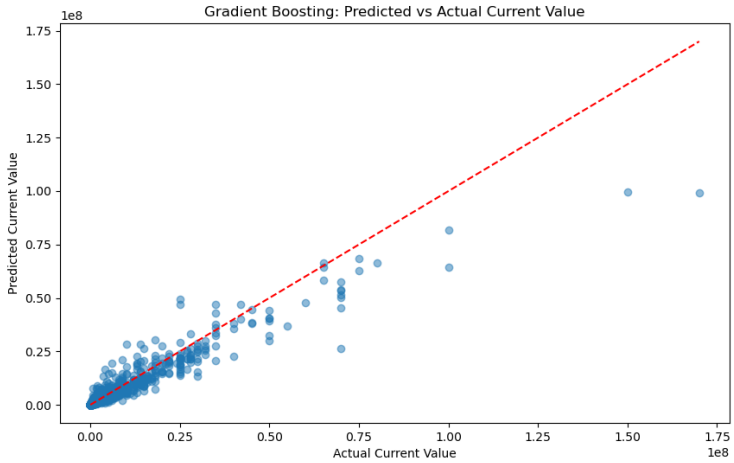

Fig. 5 Gradient boosting: predicted value compared with the actual value

A similar trend can be spotted in Figure 5, that is, gradient boosting can precisely predict the value of most low-value players while underestimating the most valuable ones on account of insufficient sample data. However, in comparison with random forest, the scatter points in Figure 5 are closer to the red line which is the real value. Hence, the gradient boosting model has a marginal advantage over random forest in this prediction.

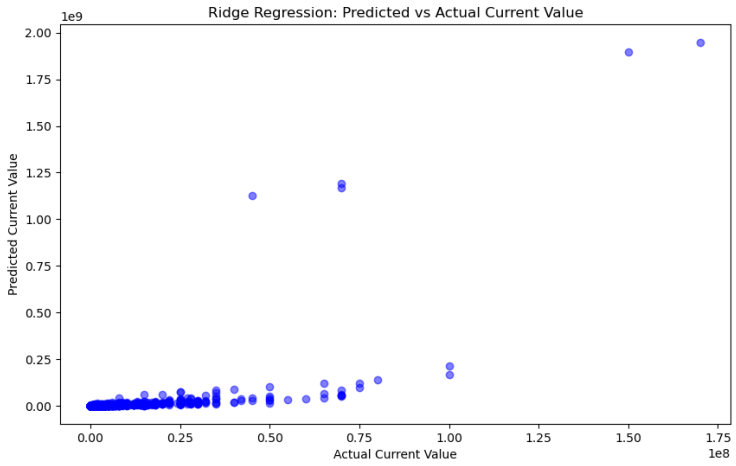

Fig. 6 Ridge regression: predicted value compared with the actual value

In Figure 6, a severe underestimation of almost all players occurs constantly. It somehow shows a barely precise prediction concerning the players whose market value is lower than around 1,500,000 euros. Since its deviation is significant, the red line indicating the actual value is not drawn. Additionally, ridge regression is not preferable in this specific prediction.

3.3. Model Evaluation

The result shows that gradient boosting achieves the best overall performance thanks to its lowest RMSE of 0.45 and highest R^2 value of 0.93, which reflects the strongest performance among these three sampled models while the random forest has a minimally better MAPE. In a word, the gap between these two is marginal, suggesting that both are appropriate for predicting the football players' current market value with relatively high precision.

As for ridge regression, due to its linearity feature, it is not working as precisely as the other two models. It may not be capable of dealing with a sophisticated dataset that contains abundant and complex variables that can hardly be correlated in linear models. However, despite its limitation in

predicting high-value footballers, ridge regression can correctly predict the players with lower current values, which form the majority of the dataset.

3.4. Feature Importance

When trying to discover the most impacting factors in random forest and gradient boosting, plots are drawn which provide a clearer visualization.

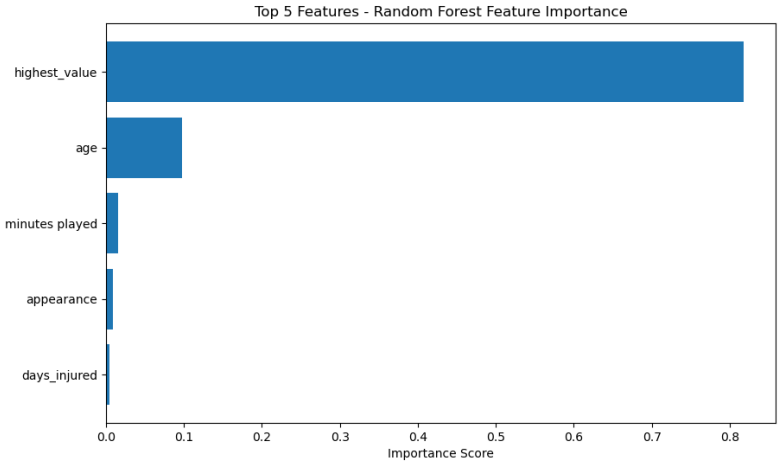

Fig. 7 Top 5 feature importance of random forest

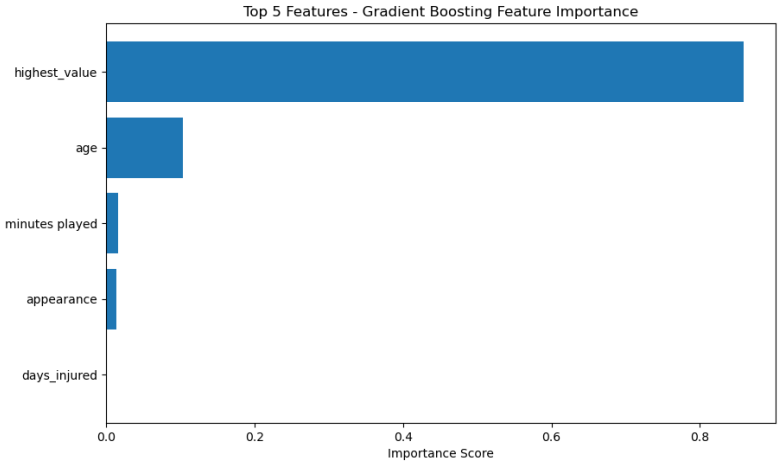

Fig. 8 Top 5 feature importance of gradient boosting

For both models, according to Figure 7 and Figure 8, the highest value is the most important factor in predicting the current value of a player, because it stands for the global achievement of a player, followed by the age and the health condition of a player, which reflect the probable potential and progress of a footballer.

4. Conclusion

In this project, three models of machine learning, random forest, gradient boosting, and ridge regression, are trained and eventually evaluated in terms of the accuracy of prediction of current values of football players. Some verification works are done in this paper. Firstly, a cross-validation process is adopted to prevent the models from overfitting and to ensure the robustness and credibility of each model. Secondly, unlike other crowd-sourced datasets, the data utilized in this project is a full-mark project shared on Kaggle. Endorsed by the high mark on Kaggle, the credibility and completeness of this database can be assured. Lastly, both linear and non-linear models are chosen in

this project, with the goal of comparing and evaluating the fitness of current value prediction. In conclusion, due to the complexity of football metrics, non-linear models for instance random forest and gradient boosting are more capable of this specific task, outperforming of ridge regression in terms of accuracy. In terms of features impacting the models, a young, robust, and healthy player with a high market value in the past is more likely to have a high current value.

In this study, the parameter of regularization of ridge regression is fixed and is relatively high, which will probably lead to the rigidness of the prediction using this model. Therefore, ridge regression's performance could potentially be improved by applying a fine-tuning method. What is more, ridge regression is not the only model of regularization, there are other models e.g. Lasso regression which may be fitter to this scenario. Additionally, the distribution of data is heavily skewed, with most players having relatively low values and only a few having very high values, this may give rise to the inaccuracy of prediction, especially for high-value players.

In the future, the improvement of this project can be commenced by applying more sophisticated modern models including XGBoost, or using a combined model in the interest of utilizing the advantage of each model. Furthermore, more variables should be included, e.g. personality or the loyalty of the player, or even the posts on social media, so that the prediction accuracy will assumably further improved.

References

[1] Poli R, Besson R, Ravenel L. Econometric approach to assessing the transfer fees and values of professional football players. Economies, 2021, 10(1): 4.

[2] Anjum S, Fatima A. Predictive Analytics For FIFA Player Prices: An ML Approach. Journal of Scientific Research and Technology, 2023: 204-212.

[3] He Y. Predicting market value of soccer players using linear modeling techniques. University of Berkeley (working paper), 2012.

[4] McHale I G, Holmes B. Estimating transfer fees of professional footballers using advanced performance metrics and machine learning. European Journal of Operational Research, 2023, 306(1): 389-399.

[5] Lee H, Tama B A, Cha M. Prediction of Football Player Value using Bayesian Ensemble Approach. Communications in Statistics-Simulation and Computation, 2022.

[6] Horning N. Random Forests: An algorithm for image classification and generation of continuous fields data sets. Proceedings of the International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences, Osaka, Japan, 2010, 911: 1-6.

[7] Firinguetti L, Kibria G, Araya R. Study of partial least squares and ridge regression methods. Communications in Statistics-Simulation and Computation, 2017, 46(8): 6631-6644.

[8] Natekin A, Knoll A. Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 2013, 7: 21.

[9] Hodson T O. Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. Geoscientific Model Development Discussions, 2022, 2022: 1-10.

[10] Chicco D, Warrens M J, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. Peer computer science, 2021, 7: e623.