International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 11 Issue 1 ǁ January 2023 ǁ PP. 149-156

Mining Analysis on the Correlation between Football Player's Competency and Value Based on Machine Learning

Mingming Shen, Siyu Wang, Mengyao Wang, Huanhuan Chen
School of Management, Tianjin University of Technology, Tianjin, 300384, China

Abstract: *With the increasing popularity of football, a complete football industry chain, represented by the five major European leagues, has gradually been formed around the world. The development of the industry chain has further driven the development of football industry in all countries. At the same time, this also means that football is no longer just a sport, and it has huge economic benefits behind it. Therefore, the players themselves and their stature have also received wide attention. Wecrawl data from FIFA's official players database, uses machine learning method, builds players' ability and value prediction model through XGBoost, and analyses the important factors that affect players' value in different positions.*
Keywords: *Machine learning; Player's stature; XGBoost algorithm*

Date of Submission: 01-01-2023 Date of acceptance: 10-01-2023

I. Introduction

Football, one of the most influential sports in the world, has been popular since its birth. Taking the 2018 Russian World Cup as an example, 3.572 billion visitors watched the event and received more than 7.5 billion people's attention, and had more than 580 million interactions in social media. Although many sponsors cancelled their cooperation due to the impact of corruption scandals, the 2018 World Cup was still sponsored by $1.45 billion, of which 21% were sponsored by Chinese sponsors. At the same time, the World Cup is broadcast to 210 countries worldwide through various platforms, and more than 500 million videos on all platforms have been watched by more than 1.2 billion people. Although some big football events have been difficult to play due to COVID-19 in recent years, the popularity of football remains high.

However, Chinese football has a tendency to stagnate or even regress. On the one hand, the training of many football players is not targeted enough. On the other hand, because the players' abilities do not match their value, and clubs choose to use funds to recruit star players at a high price rather than create a positive youth training system. Therefore, this paper will study the above two issues and achieve the following goals:

(1) Provide theoretical guidance for professionals engaged in relevant work

Many clubs and youth training schools do not have a high professional level of coaches. They often have the same training plan for their players. They continue to use the old training plans and do not tailor the training plans for players of different positions and abilities. Through the analysis of the influencing factors of players' ability in different positions, we comprehensively understand that players in different positions should have superior attributes, and better formulate special training content, so as to improve the individual competitiveness of players, improve the execution ability of team tactics system, and improve the lower limit of the whole team, so as to have a better level of performance in the face of upstream teams. At the same time, the potential value of players is predicted through the combination of the established model, which provides guidance for the team in selecting players and training reserve strength.

(2)Providing theoretical reference for the orderly development of Chinese football market

Deep capitalization has led to abnormal development of football market in China and even the world at present. Upstream players are paid much more than their own abilities, while some middle-range players are even unable to survive. Someclubs also choose to supplement their football team strength by signing star players at high prices, ignoring the importance of building a youth training system, which leads to poor training conditions and wastes players' enthusiasm. This paperwill provide theoretical reference for the clubs in evaluating players by combining their abilities with their stature, which will be conducive to the orderly development of the football market, and further allow the club to have enough funds to build up the youth team and provide sufficient backup force for the development of football.

In order to achieve the above goals, this paper takes the data of FIFA database website (https://sofifa.com) as an example. The data are taken as the object of study, through the correlation thermodynamic diagram of characteristics to analyze which features are highly related to player's value. Secondly, this paper establishes the model based on XGBoost algorithm and makes multivariate regression to discover and analyze the relationship between shortpassing, reaction, dribbling and player's value, which provides decision support for evaluating player's value and formulating training strategies.

II. Literature review

2.1 Research on the Prediction of Player Value and Match Results

Football has gradually become the largest sport in the world with its good competition atmosphere and high commercialization.With the professionalization of football leagues, players' trade and transfer is inevitable. The continuous progress of major professional football leagues makes the current transfer market have the following characteristics: on the one hand, players' value generally shows geometric growth; on the other hand, The world-famous consortium entered the football world by acquiring top clubs. According to the report of FIFA, although the world football has not gone out of the recession caused by the COVID-19, 18068 international transfers still occurred in the transfer market in 2021, with a total transfer fee of $4.86 billion. Therefore, there are a lot of researches on the player's value based on traditional statistics, and many researches on the prediction of game results and player's value based on traditional research. Chen[1] predicted the value of 275 forward players in the CSL by selecting different indicators that reflect players' ability attributes and players' influence, using correlation analysis and regression model to build a regression equation for predicting the value of registered players in the CSL. Oliveret al. [2]divided players' characteristics into three categories: basic technical attributes, on-site performance and off-site influence, and built a regression model to predict players' value. Wanbo[3] made a simple statistical analysis of the player's value and transfer fee of the CSL in 2016. It mainly studied the current situation of the domestic players' transfer market, but did not consider the impact of the players' technical ability indicators on their value. In terms of the prediction of football match results, Yanget al. [4] proposed a two-stage Bayesian model and assumed that each level in the football match would predict the match under three complete influencing factors.

2.2 Machine Learning and Football Research

With the development of computer technology, machine learning algorithm has also been applied in football research. Zhao Yan[5] uses complex network theory to build a transfer network diagram of football players, and on this basis, builds a prediction model of player's value through GBDT algorithm. The result shows that the transfer network diagram of players has a certain complementary role to the method of football player's value evaluation based on collective wisdom. Huo[6] uses wireless sensor networks to perceive players'performance on the field and record various evaluation index data, and finally uses Bayesian algorithm to build a player's value prediction model. Iman et al.[7] proposed a hybrid regression method which combines particle swarm optimization algorithm with support vector machine regression (SVR) algorithm to build a prediction model to estimate the value of transfer market players.

III. Data Acquisition

3.1 Data Explanation and Display

This paper uses Python crawlers to obtain the data of October 2022from the FIFAdatabase website(https://sofifa.com), which covers all the attributes of the player, including club, salary, personal information of the tall player, as well as advanced data such as player shortpassing, dribbing, ball control and psychological quality.

This paper integrates the player data initially crawled into a data set, including 19,980 player information and 50 characteristic indicators, of which the characteristic indicators are shown in Table 3-1

Table 3-1 Data Set Characteristics

Name	AverageValue	Standard Deviation	Meaning	Type
values	275.13998	796.58709	-	float
age	22.55371	4.199726	-	int
potential	70.49303	6.516776	-	int
reputation	1.069814	0.340055	-	int
preferredfoot	0.226077	0.41829	-	int
skillmoves	2.230936	0.735769	-	int
ST	53.30471	13.88305	shadow front position ability	int
LR_W	52.40621	15.0259	winger position ability	int
CF	52.21652	14.62246	center position ability	int
CAM	54.44101	14.36316	center position ability	int

LR_M	55.00988	14.43407	left and right midfield position ability	int
CM	53.39903	13.50768	midfield position ability	int
LR_WB	52.36555	14.13423	position ability of offensive full backs	int
CDM	51.64179	13.98353	rear lumbar position ability	int
LR_B	51.78353	14.35763	full back position ability	int
CB	50.42374	14.77773	Position ability of central defender	int
GK	22.81717	14.95748	Goalkeeper position ability	int
cross	45.37801	17.55038	-	int
finishing	43.06205	19.82093	-	int
heading_accuracy	48.40336	17.05028	-	int
shortpassing	55.15051	14.92303	-	int
volleys	39.29204	17.19778	-	int
dribbling	52.41223	19.32571	-	int
curve	43.42411	17.23775	-	int
fk_accuracy	39.19978	15.8788	free ball accuracy	int
long_passing	49.15473	15.13046	-	int
ballcontrol	54.56527	17.28765	-	int
acceleration	63.5385	15.53813	-	int
sprintspeed	63.51785	15.31501	-	int
agility	61.01505	14.85419	-	int
reactions	57.44904	9.937162	-	int
balance	63.14512	14.46144	-	int
shotpower	54.00951	13.1675	-	int
jumping	61.90642	11.40693	-	int
stamina	58.52376	16.00696	-	int
strength	60.89919	12.80057	-	int
longshots	42.62817	19.21431	-	int
aggression	50.78068	16.46289	-	int
interceptions	41.94323	20.51448	-	int
positioning	46.8581	19.61511	-	int
vision	50.577	13.42391	-	int
penalties	45.56786	15.77967	-	int
composure	52.94444	12.75005	-	int
standingtackle	44.38276	20.8988	-	int
slidingtackle	42.71705	20.25917	-	int
gk_diving	16.88054	17.78704	-	int
gk_handling	16.37526	17.00191	-	int
gk_kicking	16.20025	16.68178	-	int
gk_positioning	16.20369	16.79467	-	int
gk_reflexes	16.52028	17.74828	-	int

3.2 Data Preprocessing

Before analyzing the data set obtained by crawling, including the player's value, comprehensive ability, potential and other indicators, the data shall be preprocessed.

3.2.1 Missing Value Processing

Having reviewed and analyzed the original data set, we found that some players' data were incomplete and missing. Therefore, before analyzing the data, the incomplete part ofthe data was removed, and 17711 valid samples of data were finally obtained.

3.3 Feature Filtering and Selection

Since the data set after data preprocessing contains too many features, we will filter and select the features as followed.

The characteristics of the dataset are preliminarily analyzed and selected according to the correlation thermodynamic diagram of characteristics. The lighter the color of the thermograph is, the higher the correlation between the two characteristics is. On the contrary, the darker the color is, the lower the correlation between the two characteristics is. The thermograph of the correlation between all characteristics and player value is shown in Figure 3-3 below.

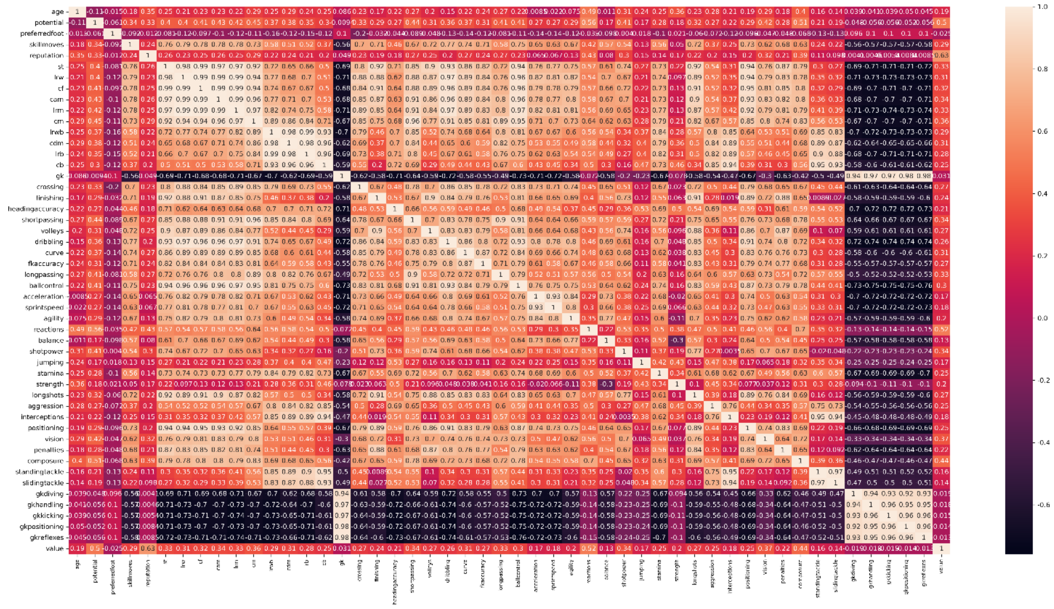


Figure 3-3the Correlation Thermodynamic Diagram of Characteristics

According to the above diagram, select some features that are highly relevant to the player's value, such as potential, reputation, shotpower and other features to establish the model.

IV. Building Player's Value Prediction Model Based on XGBoost

Based on XGBoost algorithm, a player's value prediction model is constructed. Suppose that the player's data set is X (including the player's various ability value characteristics), Y is the player's value (unit: 10000 euros), and the training data set of n samples (representing n players) and m characteristics is$D = \{(x_1, y_1), \dots, (x_n, y_n)\}$, 其中$x_i = (x_i^{(1)}, \dots, x_i^{(m)})$.

The essence of XGBoost algorithm is the Boosting iteration method, involving two key parts: the addition model (the strong evaluator is linearly added by a series of weak evaluators) and the forward distribution algorithm (the new evaluator generated in the next iteration is trained on the basis of the previous one). As an improvement of GBDT algorithm, XGBoost algorithm expands the objective function by second-order Taylor expansion, retains more information about the objective function, and adds regular terms to avoid over fitting

The processed data set is divided into training set and test set. After the data is input, XGBoost algorithm is used to train the model,predict the test set data, and evaluate the performance of the model according to the algorithm evaluation index to optimize the super parameters. The important parameters involved are max_depth (the maximum depth of the tree), learning_rate (rate of model learning), n_estimators (the number of weak estimators in the integration, that is, the number of trees), object (the type of loss function), min_child_weight (determines the sample weight sum of the minimum leaf node), regalpha (the weight of L1 regular items), reglambda (the weight of L2 regular items), subsample (the proportion of samples sampled), colsample_bytree (the proportion of features used for training each tree to all features), gamma (the minimum loss function drop value required for node splitting). Through constant parameter optimization, the best super parameter combination is: max_ depth=7 , learning_ rate=0.1 , n_ estimators=500 , object=‘reg:linear’ , min_ child_ weight=1 , regalpha=3 , reglambda=2 , subsample=0.8 , colsample_ bytree=0.9 , gamma=0.6.

The 10-fold cross-validation method is used to evaluate the results of model super parameter optimization, and the corresponding evaluation index results are shown in Table 4-1.

Table 4-1 Evaluation Index Results under 10 fold Cross validation

Evaluation Indicators	R^2	MAE	MSE
times 1	0.9999	1.4503	84.4055
times 2	0.9997	1.6641	173.0647
times 3	0.9930	2.3955	3140.7368
times 4	0.9999	1.1576	38.2470
times 5	0.9993	2.7754	354.6946
times 6	0.9994	3.0857	378.6706
times 7	0.9994	2.5500	262.9447
times 8	0.9844	5.3094	11025.6067
times 9	0.9988	3.9537	805.0503
times 10	0.9991	3.2001	478.2626
Average Value	0.9973	2.7542	1674.1683

The results show that the fitting results of XGboosst algorithm model are good.
Figure 4-2 and Table 4-2 show the Feature-importance summary diagram of the top 13 features of XGBoost algorithm respectivelyand theF-score ranking of the top 13 features.

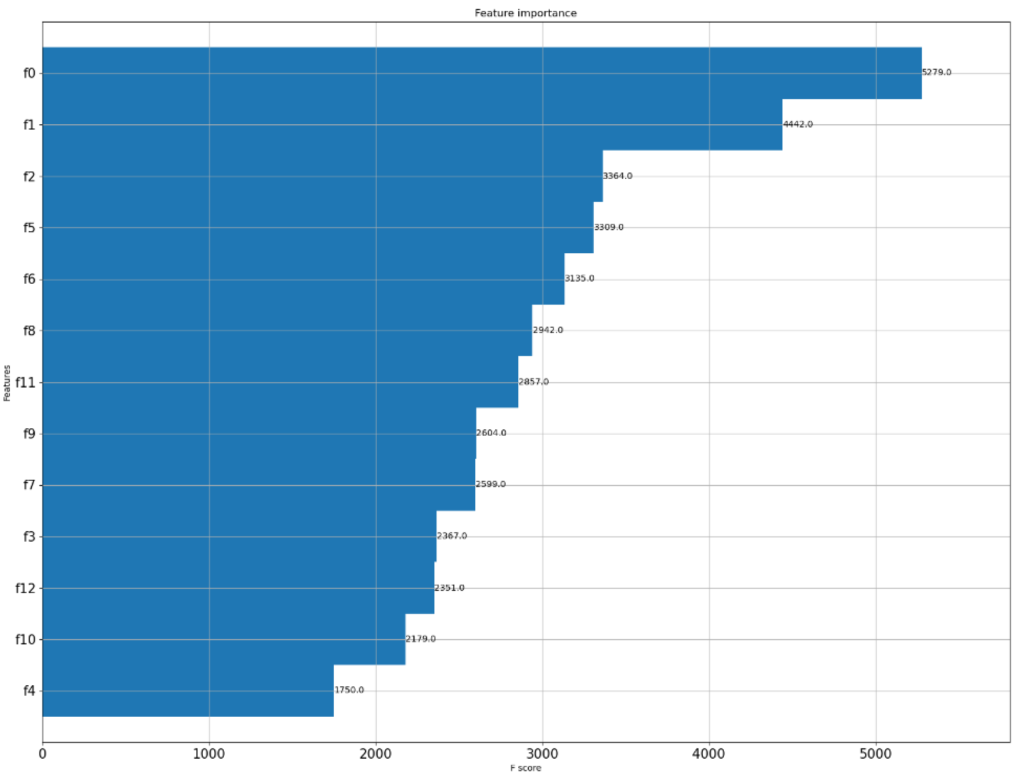


Figure 4-2 Feature-Importance Analysis of XGBoost Algorithm

Table 4-2 F-score ranking of XGBoost Algorithm

Ranking	Feature	F-score
1	age	5279
2	potential	4442
3	ST	3364
4	CB	3309
5	cross	3135
6	fkaccuracy	2942
7	reaction	2857

8	longpassing	2604
9	shortpassing	2599
10	LRW	2367
11	composure	2351
12	ballcontrol	2179
13	LRM	1750

According to Figure 4-2 and Table 4-2, age, potential, ST, CB, cross and reaction are the key factors that affect the player's value.

Specifically, age and potential are the most direct characteristics to measure a player. A young and high potential players usually has higher technical improvement space and development prospects. Through corresponding training, these young players can convert their potential into ability, and realize their talent on the court to improve their value;reaction is the basis of all the technical actions of a player on the field. It will affect a player's cooperation with his teammates in attack and defense and the ability of various technical and tactical actions; cross is a measure of a player's ability to pass the ball into the restricted area on the field. A player with a high cross value can create more chances for attacking teammates to score goals;ST,CB and other specific positionsabilityvalue also reflect the important influence of a player's ability in attack, defense and team organization on his value.

V. Analysis of Factors Affecting Player's Value in Different Positions

In the football field, players in different positions have different ability requirements and have their own contributions and tactical behaviors to the game. For example, the frontline players need to have excellent ball control skills to create shooting opportunities for teammates or score goals by themselves; the midfielder needs to have an excellent ability to read the game, and always answer the frontline teammates and backcourt teammates; backcourt players are mainly responsible for defense, interception and blocking the opponent's attack.

By analyzing the influence of the ability of players in different positions on their value, it can provide some support for evaluating the player's value and formulating training strategies.

This paper mainly constructs four data sets according to the distribution of players' positions in football matches, which are backcourt position players, midfield position players, frontline position players and goalkeeper position players.

Combined with the established XGBoostplayers'value prediction model, draw a Feature-importance summary diagram, as shown in Figure 5-1, 5-2, 5-3 and 5-4. Table 5-1 lists the F-scores of players' characteristics in different positions.

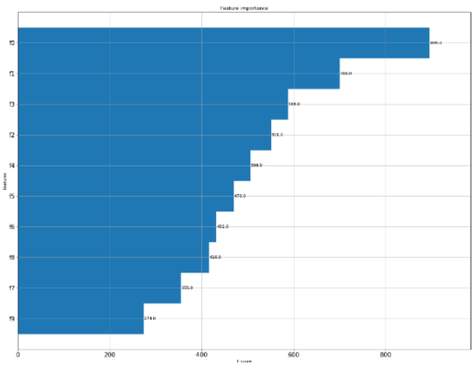
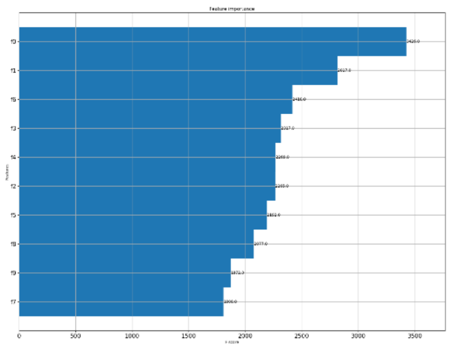
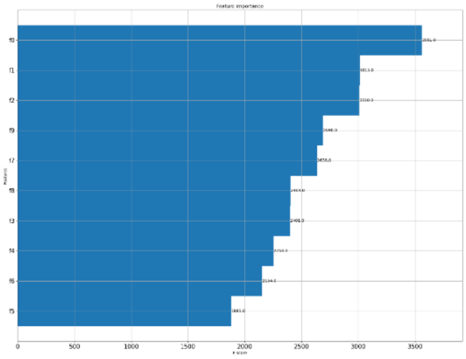



Figure 5-1 Goalkeeper Position PlayerFigure 5-2 Backcourt Position Player



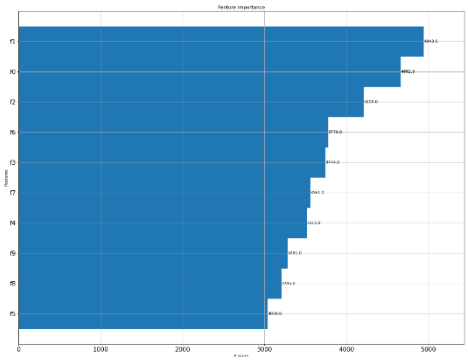


Figure 5-3 Midfield Position PlayerFigure 5-4 Frontline Position Player

Table 5-1 F-score Ranking of Characteristics of Players in Different Positions
表 5-1 不同位置球员特征重要度 F-score 排序

Ranking	Goalkeeper		Backcourt		Midfield		Frontline	
	Feature	F-score	Feature	F-score	Feature	F-score	Feature	F-score
1	age	896	age	3413	age	3561	potential	4943
2	potential	700	potential	2964	potential	3015	age	4662
3	reactions	588	reactions	2889	cross	3010	finishing	4209
4	agility	551	longpassing	2196	standingtackle	2690	agility	3778
5	composure	506	ballcontrol	2135	stamina	2638	shortpassing	3744
6	gkdiving	470	shortpassing	2122	positioning	2404	reactions	3561
7	gkhandling	432	stamina	2077	shortpassing	2401	dribbling	3515
8	gkposition	416	composure	1923	dribbling	2253	longshots	3281
9	gkkicking	355	standingtackle	1870	reactions	2154	shotpower	3205
10	gkreflexes	274	interceptions	1846	ballcontrol	1881	ballcontrol	3039

According to Figures 5-1 to 5-4 and Table 5-1, except for potential and age, the characteristics of players in different positions have different effects on their value.
(1) For the goalkeeper position players, they need to have high reactions, agility and composition, which are the basic attributes of a good goalkeeper. Whenthe opponent attacks, the goalkeeper needs to be calm, and quickly catch the opponent's shooting intention and angle, use his ability of gkposition to cooperate with his teammates in the back court, and catch the ball through his own technical and tactical ability, Then use the ability of gkkicking to accurately pass the ball to the teammates in the midfieldand frontline to create opportunities for offensive scoring. Therefore, goalkeepers need to develop reasonable training plans to improve their ability to respond, save and stand to improve their own value.
(2) For the backcourt position players, the ability of ball control, standingtackle and interceptions are important. The players in the backcourt mainly perform defensive tasks. They need to have good stamina to constantly fight, intercept and steal to undermine the opponent's attack, and through accurate longpass or shortpass,the player can pass the ball to his teammates in the midfield and frontline to turn the defense into attack. Therefore, the way for players in the backcourt position to improve their value is to promote their physical confrontation and defensive skills, and increase the accuracy of the longpass.
(3) For the players in the midfield position, the abilities of cross, positioning, shortpassing and standingtackle are more important. The midfield players mainly play the role of connecting the backcourt and the frontline teammates. They need to have a good sense of position. When the opponents attack, they can steal the ball right in time and pass the ball accurately to the frontline teammates or directly shoot into the restricted area to create scoring opportunities for the team. Therefore, the midfielders should improve their own value by raise their ability to pass in the middle and the coordination of short passes, as well as cultivating good standing ability.
(4) For players in the frontline position, the abilities of finishing, ballcontrol, shortpassing and dribbling are more important. The frontline players mainly shoot and score. After obtaining the ball right, they need to rely on their own ball control and dribbling abilities and cooperate with other frontline teammates to break through the opponent's defense in the backcourt and complete the task of shooting and scoring. Therefore, the frontline

players need to develop a reasonable plan to stimulatetheir abilities of shooting, ballcontrol and other capabilities, and strengthen their scoring ability to improve their value.

VI. Conclusion

With the gradual improvement of COVID-19, and with the convening of the 2022 Qatar World Cup, the popularity of football has risen significantly, then the player's value will still become the focus of the international transfer market. This paper constructs the football player's value model through XGBoost model, and verifies the R^2、MAE、MSE performance indicators of the model through 10-fold cross-validation, Combined with the Feature-importance of XGBoost model, this paper analyzes the factors that affect the value of players in different positions, and puts forward relevant suggestions.

Because the average field data, staged performance data and other off-site factors of players are not considered when building the model in this paper, dynamic factors can be considered in the next stage to build the model, so that the player value model is more accurate and its actual value is improved.

Reference

[1]. LIU Z, ZHOU C, CHEN H, et al. Impact of cost uncertainty on supply chain competition under different confidence levels. International Transactions in Operational Research, 2021, 28(3): 1465-1504.
[2]. OLIVER M, ALEXANDER S, MARKUS W. Beyond crowd judgments: data-driven estimation of market value in association football. European Journal of Operational Research, 2017,263(2): 611-624.
[3]. WAN BO. Study on the transfer of the super league players in winter of the 2016 Season. Bulletinof Sport Science&Technology, 2016,24(9): 107-109.
[4]. ZHOU C, TANG W, ZHAO R. Optimal consumption with reference-dependent preferences in on-the-job search and savings. Journal of Industrial and Management Optimization, 2017, 13(1): 503-527.
[5]. YANGT, SWARTZT. two-stage Bayesian model for predict winners in major league baseball.Journal of Data Science,2004,2,61-73.
[6]. ZHOU C, LENG M, LIU Z, et al. The impact of recommender systems and pricing strategies on brand competition and consumer search. Electronic Commerce Research and Applications, 2022, 53: 1-15.
[7]. LI Z, XIE H, XU G, et al.Towards purchase prediction: A transaction-based setting and a graph-based method leveraging price information. Pattern Recognition, 2021, 113: 107824.
[8]. ZHOU C, MA N, CUI X, et al. The impact of online referral on brand market strategies with consumer search and spillover effect. Soft Computing, 2020, 24(4): 2551-2565.
[9]. ZHOU C, XU G, LIU Z. Incentive contract design for internet referral services: Cost per click vs cost per sale. Kybernetes, 2020, 49(2): 601-626.
[10]. CHU M, ZHOU C, YU J. The impact of online referral services on cooperation modes between brander and platform. Journal of Industrial and Management Optimization, 2022, DOI: 10.3934/jimo.2022174.
[11]. ZHOU C, TANG W, ZHAO R. An uncertain search model for recruitment problem with enterprise performance. Journal of Intelligent Manufacturing, 2017, 28(3): 695-704.
[12]. YU J, ZHAO J, ZHOU C, et al. Strategic business mode choices for e-commerce platforms under brand competition. Journal of Theoretical and Applied Electronic Commerce Research, 2022, 17(4): 1769-1790.
[13]. HUO D. Evaluation of the value of basketball players based on wireless network and improved Bayesian algorithm.EURASIP Journal on Wireless Communications and Networking,2020, 236(9): 1-11.
[14]. ZHOU C, TANG W, ZHAO R. Optimal consumer search with prospect utility in hybrid uncertain environment. Journal of Uncertainty Analysis and Applications, 2015, 3(6): 1-20.
[15]. IMAN B, SEYED M R. A novel machine learning method for estimating football players' value in the transfer market. Soft Computing, 2020, 25(10): 2499-2511.