Dean&Francis

ISSN 2959-6157

Research on Player Market Value with Possible Influencers-Taking Attackers and Defenders as Examples

Yiwen Hua[1, *]

[1]School of Mathematical Sciences, University of Science and Technology of China, Hefei, 230026, China

* Corresponding author: Shingo_natsume@mail.ustc.edu.cn

Abstract:

This article attempts to identify those factors contributing to football player market value. To analyze and furthermore determine significant factors or independent variables with around 500 samples of football player data from 2017 to 2020, Multiple Linear Regression is the method to choose. to analyze the significant factors. Based on an assumption, 14 variables that were chosen do correlate with market value. This paper also considers the interaction effects between minutes played and the times a player in a starting lineup, and uses Forward Stepwise Regression to solve the covariance problem caused by adding interaction terms. In order to test the effectiveness of this operation, the research compares the VIF value and significance of those variables. It turns out that the number of minutes played, assists efficiency (attackers), shot on target efficiency (attackers), height(defenders) and passes frequency have a significant linear relationship with prices, while times of dribbles past, goals efficiency, times of aerial won, touches, even height (attackers) fail the significance test. Overall, the volatility of market value of attackers and defenders can be considered by the extent to which these factors affect them.

Keywords: Football player market value; multiple linear regression; interaction effects.

1. Introduction

Player market value has become an important measure in consideration when clubs are making transfer decisions. With football coming into 2020s, transfer market has received great attention. Investments into football becomes higher with wealthy merchants like Sovereign wealth fund of Qatar, the owner of Paris Saint-Germain or City Football Group (FC), new boss of Girona FC. In the past 10 years, the highest player market value has risen from 139 million Euros in 2014 (Lionel Messi) to 180 million Euros in 2024 (Kylian Mbappe & Erling Haaland & Jude Bellingham). Overall transfer fee has reached 6.4 billion dollars in the summer window of 2024, second-highest in record. Therefore, it is important to know the factors contributing to player price to make better and more financial-effective transfer choice, and aid clubs with data to find their suitable players in the fierce transfer market and competition in domestic

league or continental cups like UEFA Champions League [1].

However, factors related to player price were complicated, diverse. Rehemujiang suggested that factors like number of Appearance in International Class A competition and age can be significant. Other factors like nationality, height or birth date may have minor influences [1]. Ma pointed out that a player may reach his or her highest value at the age of 25-28, with best human motor capacity and lower frequency of injury [2]. Jose witnessed both mean and maximum market value linked to the position, team performance in leagues or age [3]. Research in Indonesia found that aging problem can influence a player's market value. Other effects like the number of minutes played or the number of times a player in the first formation will not make a difference [4].

Different indicators for players of each position on the pitch were shown in previous research by both domestic and international researchers. For strikers, possible factors include ball control, shot on targets and dribbling [5]. In midfielders, factors may change to reaction and short pass ability [6]. When it comes to goalkeepers, data match the common belief. Clearance and saves make positive contributions to a goalkeeper's market value. One research found a phenomenon in La Liga that Spanish players are more likely to have a higher market value than other players [7]. Similar correlations happen in Premier League too but with no research to support this claim.

Modern ways of evaluating or predicting market value has emerged like machine learning. Through this method, a research found the most influential factor to be the player's potential other than something in relation with performance [8]. To some extent, the high value of Mbappe or Bellingham can be explained in this way. Another research came out with a model suggesting that the market value of a player is positively influenced by age, technique, and concentration level [9]. Muller made a model in his research with factors like Log of Reddit posts, even having fixed effect with log of market value 0.026 [10].

This article will use the multiple linear regression model to find correlation between statistical numbers (Age, Height, Games played, League, Goals and Assists, Passes, Shots on target, Tackles, Dribbles past, Aerials won, Clearnce) and market value shown on the professional website Transfer market. In order to get an unbiased result and eliminate the influence of injuries, the research will use statistical number per 90 mins instead of data for all season. The ultimate goal is to determine every factor with its contributions to players market value.

2. Methods

2.1 Data Source

The dataset used in this paper is fetched from the Kaggle website (Soccer players values and their statistics). All data are collected from Season 2017-18 to 2019-20, within the 5 main League of Europe (Premier League in England, Series A in Italy, La Liga in Spain, Bundesliga in Germany, Ligue 1 in France). This dataset has around 2.8 million groups of data for 6000 players (a player with 3 seasons is counted thrice). But only around1000 of player data will be in use.

2.2 Variable Selection

The original dataset boasts a large amount of data, and different players' statistics on one measure vary. For instance, no attackers will have any success of saves because they are literally not goalkeepers. So this research will only focus on two common groups, attackers (including strikers, wingers and Central Forwards) and defenders (including left backs, right backs and central backs). Eventually, a random sampling is done to get 500 observations respectively on 2 positions. The general indicators consist of 6 variables (Age, Height, League, Minutes played, Touches). For Attackers, this article focuses on Goals, Assists, Shots on Target, Dribbles past. Defenders will be measured and evaluated on Passes, Clearance, Tackles won, Aerials won. Table 1 gives a detailed description:

Table 1. List of Variables

Variable	Logogram	Meaning
Age	x_1	The age of a player
Height	x_2	The height of a player
League	x_3	Premier League (1), Series A(2), La Liga (3), Bundesliga (4) and Ligue 1(5)
Starts	x_4	A player plays a match with himself be in the starting lineup
Minutes Played	x_5	Minutes played for a player in a regular season (all games counted)

ISSN 2959-6157

Touches	x_6	Number of times when a player touches a ball
Goals	x_7	Number of goals
Assists	x_8	Number of assists
Shots on target	x_9	Number of shots on target
Dribbles past	x_{10}	A player takes the ball to pass a opponent
Passes	x_{11}	Number of times when a player passes a ball
Tackles won	x_{12}	A player made a tackle to win the ball back
Clearance	x_{13}	A player kicks a ball away to release pressure
Aerials won	x_{14}	A player jumps over to touch the ball with head while opponents around
Market Value	Y	Player's Value judged by professionals in economics

2.3 Method Introduction

The paper uses a multiple linear regression model to compare the situation with and without considering the interaction terms. This section will mainly aim to compare the significance of the two models and the accuracy of the results. Eventually, it will enable the optimized processing of models.

The multiple linear regression model is a linear regression model with multiple explanatory variables. It is a tool to check possible linear relationship between the explained variable and multiple other explanatory variables. Moreover, its basic principle or working method is to estimate a set of parameters by Ordinary Least Square (OLS) targeting at minimizing the sum of all residuals between the dependent variables and independent variables squared.

3. Results and Discussion

3.1 Multiple Linear Regression

The analysis in this paper shows that there are many factors influencing a player's market prices. As Figure 1 and 2 show:

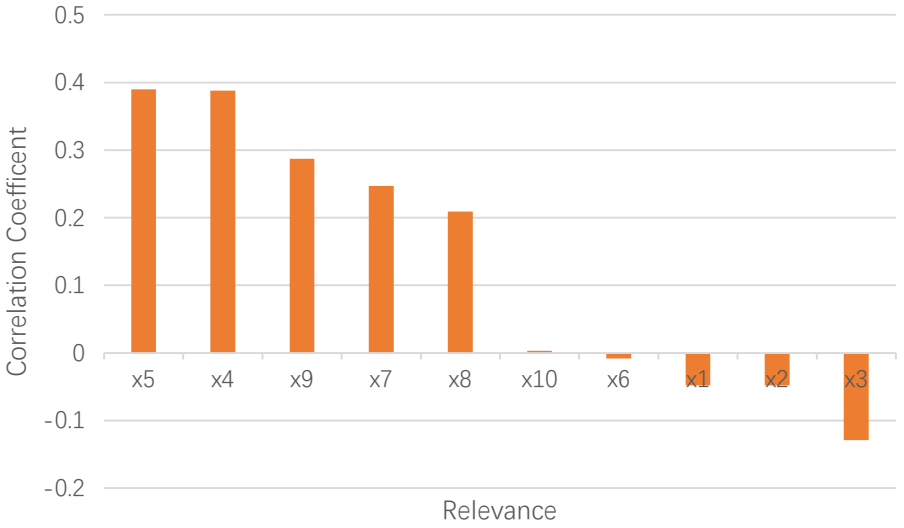

Fig. 1 Relevance Analysis Between Dependent and Independent Variables (Attackers)

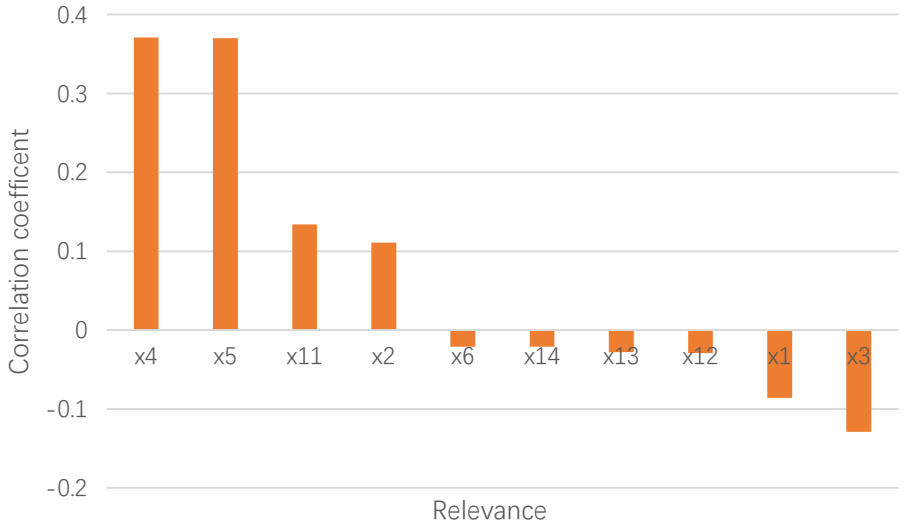

Fig. 2 Relevance Analysis Between Dependent and Independent Variables (Defenders)

Figure 1 and 2 present Pearson correlation coefficient between these factors and market value. The research data found that the number of game starts, minutes played, and shots on target per 90 mins are factors that correlate most positively with market value of attackers. This corresponds with the target for attackers: play on the ground to make goals. For defenders, both game starts and minutes played are also key factors, along with passes per 90 mins, which means the ability to pass balls are now of great importance than making clearance and easing threats. Goals and assists per 90 mins for attackers, or height for defenders are also positive correlation factors not as significant as the factors above. As for league factor, better league (no one will consider Ligue 1 better than Premier League) does indicate higher market value. From all the above, what affect the market values for players are comprehensive. Clubs now are saving their money to trying to get players they wanted. After analyzing the Pearson correlation matrix of various factors, multiple regression analysis was conducted. The general mathematical model for multiple linear regression is:

$$E(Y_1) = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_6 x_6 + \alpha_7 x_7 + \ldots + \alpha_{10} x_{10} + e_1 \quad (1)$$

$$E(Y_2) = \beta_0 + \beta_1 x_1 + \ldots + \beta_6 x_6 + \beta_7 x_{11} + \ldots + \beta_{10} x_{14} + e_2 \quad (2)$$

Here, Y_1 represents market value of attackers ($1 unit = 10,000 euros$), while Y_2 is the market value of defenders. In the above formula: α_0 and β_0 are two constant terms, e_1, e_2 are residual terms.

Table 2. Regression coefficient table for Y_1

	B	S.E.	Beta	T	significance	VIF
Const.	4096.946	1914.600	-	2.140	0.033	-
X_1	-70.759	16.379	-0.189	-4.320	0.000	1.107
X_2	-12.237	10.338	-0.050	-1.184	0.237	1.039
X_3	-134.425	48.325	-0.117	-2.782	0.006	1.023
X_4	43.287	29.522	0.296	1.466	0.143	23.719
X_5	0.137	0.351	0.080	0.390	0.697	24.137
X_6	0.963	2.459	0.017	0.392	0.696	1.151
X_7	424.871	226.124	0.095	1.879	0.061	1.500
X_8	1079.249	448.323	0.104	2.407	0.016	1.092
X_9	471.801	128.484	0.187	3.672	0.000	1.513
X_{10}	23.182	21.126	0.048	1.097	0.273	1.135

ISSN 2959-6157

Table 3. Regression coefficient table for Y_2

	B	S.E.	Beta	T	significance	VIF
Const	-2104.262	1728.277	-	-1.218	0.224	-
X_1	-51.959	11.590	-0.184	-4.483	0.000	1.090
X_2	21.024	9.282	0.102	2.265	0.024	1.320
X_3	-114.605	34.057	-0.136	-3.365	0.001	1.055
X_4	-47.423	38.808	-0.436	-1.222	0.222	82.417
X_5	1.035	0.440	0.838	2.352	0.019	82.409
X_6	0.070	0.420	0.107	0.167	0.867	267.661
X_{11}	4.380	1.168	0.184	3.748	0.000	1.562
X_{12}	-74.777	87.929	-0.034	-0.850	0.395	1.044
X_{13}	-28.438	23.702	-0.056	-1.200	0.231	1.413
X_{14}	-4.500	29.772	-0.097	-0.151	0.880	266.788

Table 2 and 3 shows the regression coefficients of the multiple linear regression equation model. For Y_1, The p-values of the T-test for the two independent variables x_1 and x_9 did not exceed 0.003. For Y_2, the p-values of x_1, x_4 and x_{11} are lesser than 0.003. These independent variables can be classified as influential of the dependent variable Y_1 or Y_2. Based on the data above, the relevant multiple linear regression equations are revealed:

$$E(Y_1) = 4096.946 - 70.759x_1 - 12.237x_2 + \ldots + 23.182x_{10} \quad (3)$$

and

$$E(Y_2) = -2104.262 - 51.959x_1 + 21.024x_2 + \ldots - 4.5003x_{14} \quad (4)$$

The coefficient R2 for fitting multiple linear regression are 0.270, 0.222, and the adjusted R-squared is 0.253, 0.206. The models have a good fit.

3.2 Multiple Linear Regression (MLR) with Interaction Terms

Interactions between some independent variables may also have some effect on market value, and these terms with interactive effects are called interaction terms. For example, a player can only make an attempt to pass a ball or do a dribbing while he touches the ball once, i.e., the influence of x_{10} and x_{11} on $y_1(y_2)$ are strongly influenced by x_6.Moreover, players with more minutes to play on the pitch are more likely to be selected to be on the starting lineups, i.e, the influence of x_4 on y_1 and y_2 is strongly influenced by x_5:

$$E(Y_1) = \alpha_0 + \alpha_1x_1 + \ldots + \alpha_6x_6 + \alpha_7x_7 + \ldots + \alpha_{10}x_{10} + \alpha_{11}x_4x_5 + \alpha_{12}x_6x_{10} + e_1 \quad (5)$$

$$E(Y_2) = \beta_0 + \beta_1x_1 + \ldots + \beta_6x_6 + \beta_7x_{11} + \ldots + \beta_{10}x_{14} + \beta_{11}x_4x_5 + \beta_{12}x_6x_{11} + e_2 \quad (6)$$

Where α_i and β_i ($i = 1,2,\ldots,12$) are regression coefficients, x_4x_5, x_6x_{10} and x_6x_{11} are interaction terms.

If the interaction term regression coefficients are significantly positive, it indicates that more minutes played on the pitch are expected to result in higher market value when this player may have more times to start a match. With the addition of the interaction term, the significance of the original regression coefficients of the independent variables becomes less important than it was originally. The regression coefficient for x_5 is $\alpha_5 + \alpha_{11}x_4$ and $\beta_5 + \beta_{11}x_4$ respectively, so the significance of α_5 or β_5 alone do not reflect whether the overall effect of x_5 on y is significant.The situation is exactly the same with x_6. The results of the analysis using the multiple linear regression model are shown in the table 4 and 5 below:

Table 4. MLR Model analysis results with interaction terms (Y_1)

Variables	α_i	Coefficient	T Value	P Value	VIF	Tolerance
Constant	3751.595	-	1.953	0.051	-	-

x_1	-73.765	-0.048	-4.437	0.000**	1.145	0.873
x_2	-10.095	-0.048	-0.972	0.332	1.052	0.951
x_3	-144.731	-0.118	-2.969	0.003**	1.045	0.957
x_4	17.129	0.388	0.504	0.614	31.511	0.032
x_5	0.496	0.390	1.186	0.236	34.299	0.029
x_6	-0.720	-0.008	-0.273	0.785	1.328	0.753
x_7	425.665	0.247	1.886	0.060	1.500	0.667
x_8	1009.125	0.209	2.241	0.026	1.106	0.904
x_9	482.591	0.287	3.759	0.000**	1.516	0.659
x_{10}	20.614	0.003	0.973	0.331	1.146	0.873
x_4x_5	-0.001	0.379	-0.171	0.864	11.499	0.087
x_6x_7	0.075	0.023	1.871	0.062	1.531	0.653

Note: ** denotes $p < 0.01$, which shows a significant effect.

Table 5. MLR Model analysis results with interaction terms (Y_2)

Variables	β_i	Coefficient	T Value	P	VIF	Tolerance
Constant	-2446.592	-	-1.418	0.157	-	-
x_1	-50.260	-0.086	-4.315	0.000**	1.116	0.896
x_2	21.574	0.111	2.326	0.020	1.336	0.749
x_3	-121.984	-0.138	-3.592	0.000**	1.063	0.940
x_4	-65.846	0.371	-1.664	0.097	86.895	0.012
x_5	1.079	0.370	2.405	0.017	86.866	0.012
x_6	5.604	-0.021	2.922	0.004**	5647.855	0.000
x_{11}	8.740	0.134	4.669	0.000**	4.063	0.246
x_{12}	-99.783	-0.029	-1.110	0.268	1.106	0.904
x_{13}	-25.667	-0.028	-1.077	0.282	1.449	0.690
x_{14}	-21.569	-0.021	-0.716	0.474	276.967	0.004
x_4x_5	0.004	0.365	0.820	0.412	15.117	0.066
x_6x_{11}	-0.059	-0.020	-2.949	0.003**	4963.026	0.000

There are many variables with *VIF* values greater than 5 and low correlation coefficients, which indicates that the inclusion of interaction terms leads to severe covariance problems in x_4, x_5, x_6, x_{11}, x_7, x_4x_5, x_6x_{11}. To solve this problem, Forward Stepwise Regression is used. There is the obtained from the regression analysis of the independent and dependent variables (Table 6 and 7):

ISSN 2959-6157

Table 6. Results of Forward Stepwise Regression (Y_1)

Variables	α_i	Coefficient	T Value	P Value	VIF	Tolerance
Constant	1977.716	-	4.616	0.000**	-	-
x_1	-72.793	-0.048	-4.451	0.000**	1.095	0.913
x_3	-121.570	-0.118	-2.525	0.012*	1.008	0.992
x_5	0.657	0.390	8.525	0.000**	1.153	0.867
x_8	971.022	0.209	2.171	0.031*	1.078	0.927
x_9	574.078	0.287	5.331	0.000**	1.054	0.949

Table 7. Results of Forward Stepwise Regression (Y_2)

Variables	β_i	Coefficient	T Value	P Value	VIF	Tolerance
Constant	-1303.025	-	-0.850	0.396	-	-
x_1	-53.886	-0.086	-4.708	0.000**	6.165	0.162
x_2	16.137	0.111	1.987	0.047*	1.719	0.582
x_3	-119.197	-0.138	-3.537	0.000**	2.192	0.456
x_5	0.498	0.370	10.026	0.000**	1.541	0.649
x_{11}	3.794	0.134	2.965	0.000**	1.021	0.979

Note: ** denotes p<0.01, which shows a significant effect.

The models identify x_1, x_3, x_5, x_8 and x_9 as explaining 25.6% of the change in y_1, and x_1, x_2, x_3, x_5 and x_{11} as explaining 21.5% of the change in y_2. Both models successfully pass the F-test ($F_1 = 29.504, F_2 = 27.793 P_1 = P_2 = 0.000 < 0.05$), which indicates that models are valid. Then the model formulas are:

$$y_1 = 1977.716 - 72.793x_1 - 121.570x_3 + 0.657x_5 + 971.022x_8 + 574.078x_9 \quad (7)$$

$$y_2 = -1303.025 - 53.886x_1 + 16.137x_2 - 119.197x_3 + 0.498x_5 + 3.974x_{11} \quad (8)$$

4. Conclusion

The study selected 435 samples of attackers and 516 sample of defenders from 2017 to 2020 from the data set, 14 variables under research. The method chosen (Multiple linear regression analysis) is accurate, effective, and comprehensive. It performs a multifactor analysis. The Pearson correlation coefficients of each variable are calculated and found.

During the analysis stage, the article uses a multiple linear regression model in order to evaluate the possible relationship between the variables and market value. For more accurate and detailed results, interaction effects are taken into account and interaction terms added with coefficients to the equation. As for results, the factors owning a positive coefficient with player market value are the number of minutes played, assists efficiency (attackers), shot on target efficiency (attackers), height(defenders) and passes frequency. Age is a factor with negative impact on market value, however. The most important factors are minutes played and the frequency for certain motions.

With the research, clubs searching for good players in the market obtains more angles in decision-making, and then have an overall determination on the budget of transfer fee. However, the drawbacks are that causal relationships between variables can't be found, the sample size need to be enlarged to maybe other leagues in small countries, and the data collected is already out of date because Year 2017 have been something 7 years ago. Also, Season 19-20 is a complicated season due to the global COVID-19, and after that many clubs cut down their expenditures, changed

their strategy of transfer window and even sell players to pass the Free Play Policy, also known as FFP. To improve this, searching for latest data (maybe Season 2022-23) and using the control variable method to find out possible causalities between player market value and factors are possible and practical ways.

References

[1] Rehemujiang. Repukati. Research on the Correlation between the Relative Age Effect of Football Players and Their Value – Take 39 national men's youth foorball teams as an example. Beijing Sport University, 2019.

[2] Ma Rui. Research on Professional Football Players' Value and The Influencing Factors. Tianjin University of Finance and Economics, 2020.

[3] Jose Luis Felipe, Alvaro Fernandez-Luna, Pablo Burillo, Luis Eduardo de la Riva, et al. Money Talks: Team Variables and Player Positions that Most Influence the Market Value of Professional Male Footballers in Europe. Sustainability, 2020, 12(9): 3709.

[4] Hajar Iman Adiwiyana, Iman Harymawan. Factors that Determine the Market Value of Professional Football Players in Indonesia. Jurnal Dinamika Akuntansi, 2015, 13(1): 51-61.

[5] Boden B P, Dean G S, Feagin J A, et al. Mechanisms of anterior cruciate ligament injury in basketball: video analysis of 39 cases. Orthopedics, 2000, 23(6): 573-578.

[6] Liao Bing, Wang Zhi-ning, Li Min, Sun Rui-na. Integrating XGBoost and SHAP Model for Football Player Value Prediction and Characteristic Analysis. Computer Science, 2022, 49(12).

[7] Shounak Sengupta. Understanding La Liga: Are match performances and player market value related? International Journal of Advanced Research, 2021, 9(01): 12-21.

[8] Mustafa A. Al-Asadi, Sakir Tasdemır. Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques. IEEE Access, 2022, 22631-22645.

[9] Ahmet Talha Yiğit, Barış Samak, and Tolga Kaya. Football Player Value Assessment Using Machine Learning Techniques. INFUS, 2020, 10: 289-297.

[10] Oliver Müller, Alexander Simons, Markus Weinmann. Beyond crowd judgments: Data-driven estimation of market value in association football. European Journal of Operational Research, 2017, 263: 611-624.