A Transfer Value Approach to Analyse Football Players

Rashmi Gandhia[a], Sudhanshu Ranjan[b]*, Nikunj Agarwal[c], Anupam Sharma[d]

[a,b,c,d] *CSE, ASET, Amity University U. P. Noida, India*

Abstract

The world of football transfers has often troubled experts of both football and economics alike. Player valuations and their transfer fees are determined opaquely and the factors that goes into determining them are highly irregular and sporadic. In this work, factors are computed that determine what transfer fee a player is worth. Also found out if these factors limited to on-field metrics, what these metrics are and the extent of their influence. The dealings of football transfers have always been seen as being ambiguous, unpredictable, and uncharacterizable. This work aims to find whether these suppositions are indeed true or are definitive factors that determine a player's worth. As investigation goes with the accuracy with which we can estimate a player's present valuation (a player's market worth), as of September 1st 2021, machine learning algorithm to provide answers to these concerns. Here, metrics identify best translate into a higher transfer value for players. Six regression models are tested and do prediction the transfer values of players based on the aforementioned metrics.

Keywords: Machine learning; Regression; Football Transfers; Sports Econometrics

1. Introduction

As we all know that how the sports industry is churning a large amount of money nowadays & Football is the biggest shark in that money pool but what about the players who play these sports. How does that value of each player taken out, like on what basis. In this work, net value of the player deserves and how much valuable is he considering all the factors which are there to be looked upon [1-4].

Transfers are an integral part of the club football, where the players are transferred from one team to another, where the biggest of clubs exchanges their player for an enormous amount of money to sign the player which they want the most. In 2019, It was revealed by the world's biggest football organising body FIFA that the clubs spent a total of $7.30billion in the transfer market [7,8]. It is like a business transaction which happens between two different clubs. It doesn't matter if the club doesn't even play in the same league or in the same country where a player moves from one to another club according to the deal. The transfer fee differs for each player it does depend on many factors like commercial value, potential worth if the player in the future, quality if the player, contract time left with current club, performance of the player, etc.

A transfer happens only when two clubs mutually agree on the term of selling the player where the club who is buying the player creates and agrees with the new contract on which they are signing the player. It can happen when a representative from a club makes a move by making an official inquiry for the target player to the club. If the club is open to sell the player, then it is obvious that the other clubs might get involved. There are many major factors which are considered while the transfer process it is discussed with their agents and advisors. To understand it by an example like how Manchester city completed Haaland transfer. E. Haaland the Norwegian moved from the German club Borussia Dortmund in summer 2022 for $63million to the English premier league's powerhouse Manchester city where he signed a five-year deal.

2. Literature Review

The sports industry is churning a large amount of money nowadays & Football is the biggest shark in that money pool but what about the players who play these sports. How does that value of each player taken out, like on what basis and a significant amount of research has been done in this field of work.

Researchers like Carmichael & Thomas (1993) had adopted an approach called the bargaining approach. In the player transfer market, it seems there are two factors that determine transfer fees. A player's value to the selling and buying clubs can be seen on the one hand, while on the other there is the position of both the selling and the buying side in the negotiations or bargaining. Using multilevel regression techniques, the authors of determine the market values of the payers [1-3].

* Sudhanshu Ranjan
E-mail address: sudhanshuranjansr1@gmail.com

Carmichael et al. (1999) consider that intrinsic talent and expenditures in human capital are what primarily determine a player's addition to the value of the team [1]. Pujol et al evaluate collegiate athletes by a ‘media value measure’[4]. Frick also includes studies looking at the quantity of Google hits to determine whether there is a "superstar impact" [5]. The impact of estimates made by the football audience on the player's transfer fee is covered [6]. The market value of football players as estimated by professional football fans is used by the authors to estimate the players' transfer fees. Using machine learning approaches, J. Hucaljuk and A. Rakipovi seek to forecast football scores [7].

According to E. Franck and S. Nüesch, reputation and ability both affect a football star's market worth. Twenty criteria were utilised to measure a player's ability in relation to his capacity to advance his team's fortunes. The likelihood that a player may sustain an injury is another element that directly affects the number of games played and, as a result, has an impact on player performance. This statistic has not, as far as we are aware, appeared in empirical study mainly because of scarcity of data [8-9].

Meanwhile, studies have shown that players who play or are approached by the more economically sound leagues are usually valued higher. Furthermore, players that play in UEFA competitions are also usually valued higher than their counterparts, although they command higher [10] salaries and even if their on-field performances are alike [11]. Van Den Berg expresses similar conclusions, that clubs with a larger buying size viz. clubs participating in UEFA competitions, pay larger values for players [12]. They also not attributed to said clubs simply having a larger purchasing power, but to a phenomenon van den berg expresses as ‘risk-aversion’ which is clubs buying players before they join their rivals.

According to McHale et al. however, the economic factors are still somewhat secondary when it comes to valuation of players in the market [13]. Their study suggest transfer values for players are still determined in a very rudimentary way, based on on-field performances. They however outline, not all teams succeed in the correct valuation of players and hence there are plenty of cases of clubs over-valuing on undervaluing players. Clubs, however, seemingly follow patterns in this process of undervaluing/overvaluing players and in [13], it is presented clearly that teams that succeed in correctly valuating a player’s on-field attributes, succeed repeatedly in making smart purchases.

3. Methodology

3.1 Data Gathering

For obtaining the data for this research two sources which were reliable that are FBREF and Transfermarkt are used.

3.1.1 Player Performance Data

Measurement of football players performance on the field (competing in the top stages around the world) are tracked by data at fbref.com. The website's data collection began with the 2017–18 season, when it became the most complete. Prior to that, there were either fewer measures or just no data available. To maintain consistency, it was concluded that it would be appropriate to use data beginning in 2017 and ending in 2021.The EPL, La Liga, Serie A, Bundesliga, Uber Eats Ligue 1 had the most complete data, while the website does provide player performance data from other leagues across the world. As a result, it was determined that information from the preceding four seasons should be obtained for each of the top 5 leagues. We used player’s statistics from prior years as we wanted to examine the idea that these metrics from previous years will significantly affect their transfer values. After collecting the data, the 140 downloaded spreadsheets were systematically combined into one.

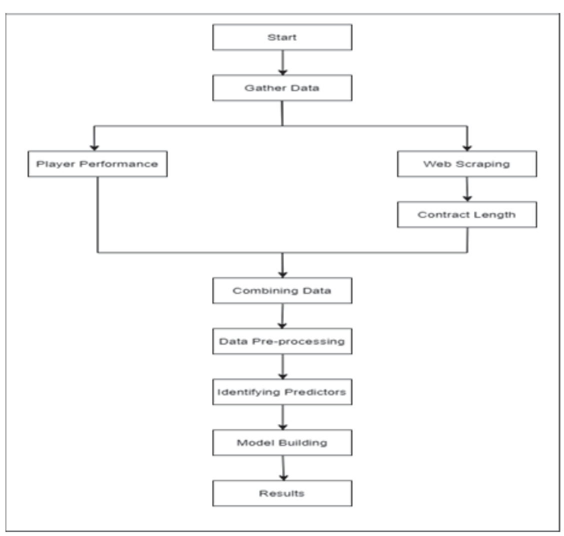

Fig 3.1 Workflow Methodology

Goalkeepers were excluded from this project due to the fact that they are evaluated using entirely different measures than outfield players. We disregarded the dataset pertaining to the performance and behaviour of the goalkeeper. A summary of the other data is as following-: Basic Stats, Finishing, Vision, Pass Types, Creativity, Defensive Actions, Possession & Other Information like additional dat for example quantity fouls they committed & drawn themselves, offside etc[14].

This marked the end of the data gathering and processing for our 1st source. Our 2nd source that is Transfermarkt.co.uk provides data about their name, group they are currently, homegrown association in which player's group contends, age, their playing position, number of years left in their agreement and the value of their transfer.

3.1.2 Web Scraping for Additional Data

The data on this website was extracted through the use of the widely known Python scraping tool BeautifulSoup. In order to gather information from all 98 different teams in the top five European leagues, we employed the 'find all()' method to obtain links to the league webpages, as compiling a list of links for each of the 98 websites would have been a time-consuming task.

Once we had identified the specific data we wished to extract, we utilized the find all() method to scan the webpages of each team in order to extract information on the contract length of individual players. This information is critical in determining a player's value. We achieved this by creating a loop to compile a list of links for every player's page on each club's website. Thankfully, we were able to extract the contract expiration dates for each player by using the find all() method on their respective webpages. Finally, we combined the contract length data with the other extracted data using the pd.merge technique. Our task of gathering data from Transfermarkt.co.uk is now complete.

3.2 *Combining FBREF and Transfermarkt data*

After completing the cleaning phase, we merged all the collected data into a single dataframe. Players who did not compete in a particular season were identified by the NaN values in their statistics metric columns. We did not want to discard the acquired data due to the presence of NaN values because such rows are excluded when fitting data into a model. To replace the NaN values with appropriate figures, we developed custom code that computes the mean statistics metrics of each action for seasons in which the player participated in the top 5 leagues.

We did not want the NaN values for a player to depend on the statistics of other players, so we avoided using a straightforward fillna() function with column averages. Instead, we created a function that replaces missing data with the mean of a player's own recorded statistics. Since our dataset had an insufficient number of goalkeeper columns to train an AI model, we excluded all goalkeepers from the final dataset with no NaN values. Therefore, goalkeepers are not considered in our study.

3.3 Data Pre-Processing

Now that we had a clean dataset that had been combined, we separated to the dataset to include only attacking players and analysed the distribution of specific predictor features that have the potential to predict the transfer values of attackers. Fig. (3.3) demonstrates this.

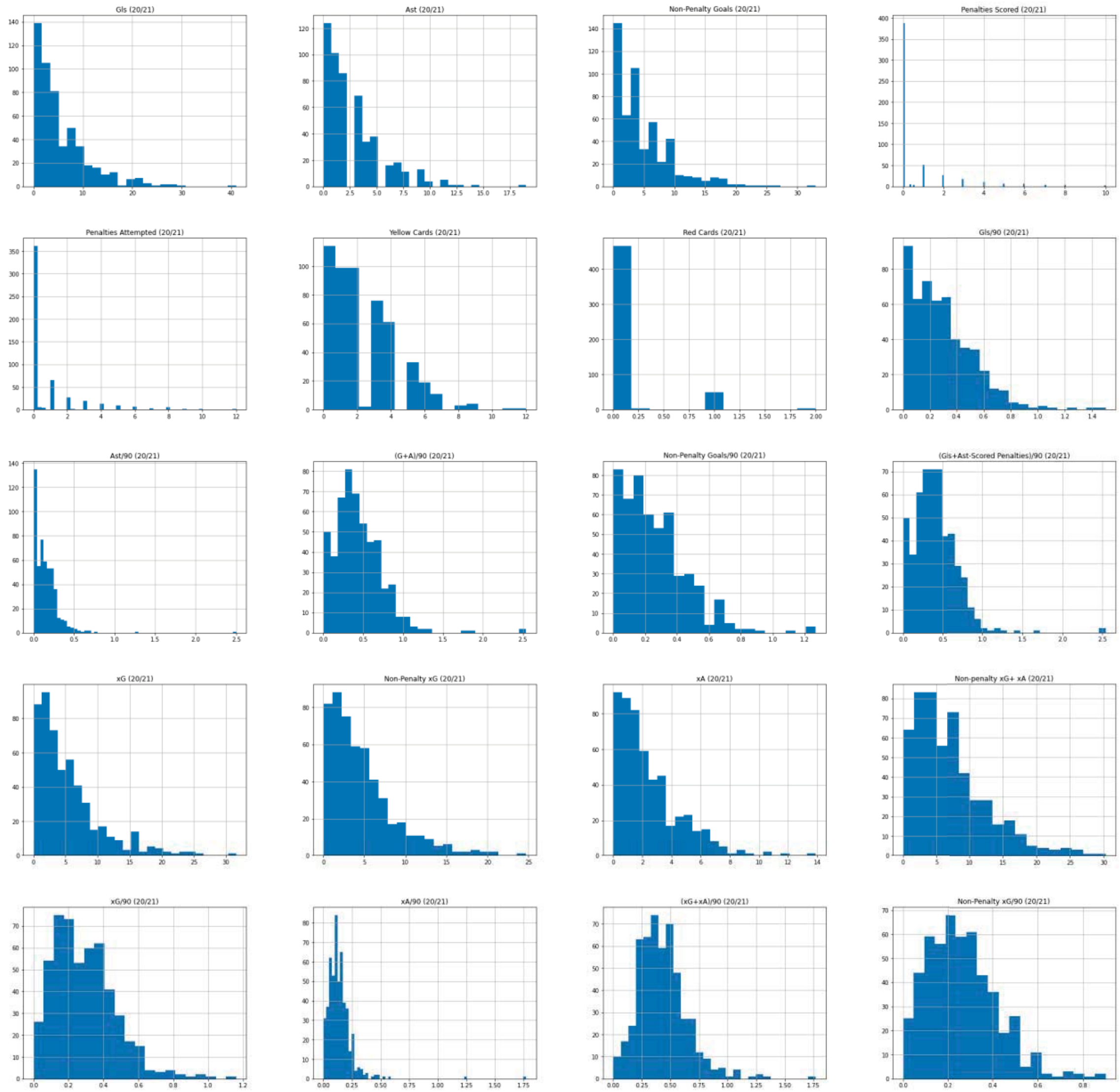

Fig. 3.3 Predictors for attackers

From our study we observed that there are probably key indicators of how valuable an attacking player are not evenly spread out in their dissemination of values. It demonstrates that before these metrics can be utilised to fit and train our models, they needed transformation to obtain a more Gaussian distribution. This was achieved using the Power Transformer from the sklearn pre-processing toolkit. We carried out standardization using a Robust Scaler. The Robust Scaler was used because it is very helpful when working with data that contain outliers.

3.4 Identifying Top Predictors for each position

The following images show some of the key characteristics that correlate to the transfer sum paid for players of all the three areas of the pitch we're observing. Each model's most crucial features were identified using the 'feature importance ' property after the data had been initially fitted into it.

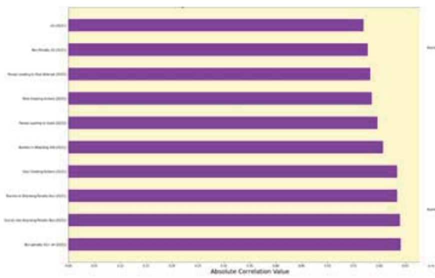

Fig. 3.4.1 Attackers

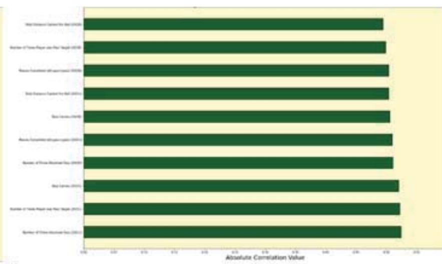

Fig. 3.4.2 Midfielders

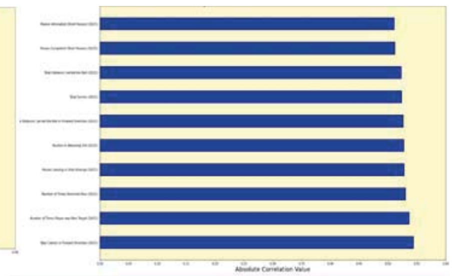

Fig. 3.4.3 Defenders

Fig. 3.4 Top predictors for different Positions

As expected for forwards from Fig. (3.4.1) a player's xG and xA, actions in the opponent's 18-yard box area and ability to create chances are the most important attributes. Similarly, in the case of midfielders Fig (3.4.2), it is no shock that they're valued based on how well they dictate the pace of play and positively assist the team's ball movement forwards. It's noteworthy to note that among the top predictors for defenders Fig (3.4.3), the quantity of tackles and interceptions made does not rank highly. Top defenders may not need to make as many tackles or interceptions since they play for stronger teams that have the ball for longer stretches of time.

3.5 Model Building

To forecast the transfer values of the players for this project, 6 different models were applied. Data had been pre-processed for each of them using the aforementioned procedures. All of our models underwent pre-processing using the aforementioned steps, and the default hyperparameters were used for the initial attempts. The entire dataset was utilized during these initial modeling attempts to determine the most important features for each model. After identifying the top 10 features, a subsequent iteration of the same model was tested using a dataset that only contained the most influential attributes as shown in Fig. (3.5). We only used the top features in the dataset for our final models. Then, after a Grid Search revealed better hyperparameters, a fresh model fitting was carried out using the newly discovered hyperparameters as well as the dataset's top attributes. Following this, a scoring metric "neg root mean squared error" was used to obtain the mean cross-validation scores.

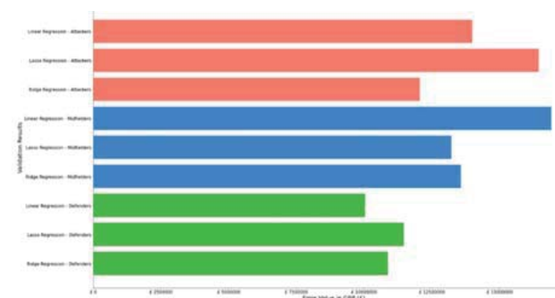

Fig. 3.5.1 Linear Regression

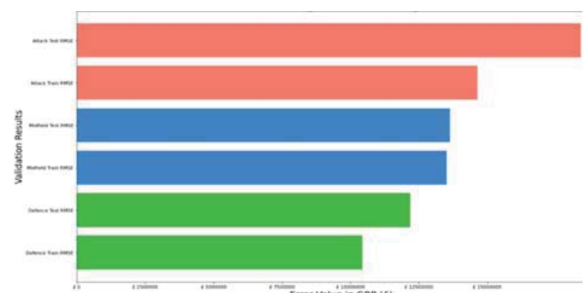

Fig. 3.5.2 Decision Tree Regressor

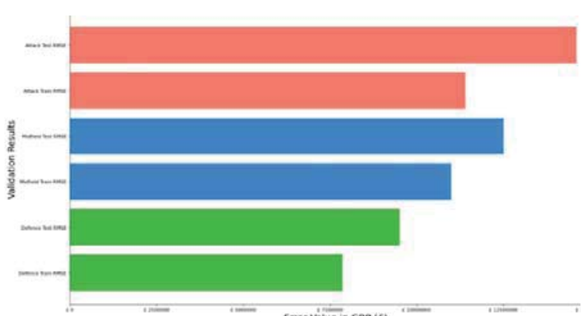

Fig. 3.5.3 Random Forest Regressor

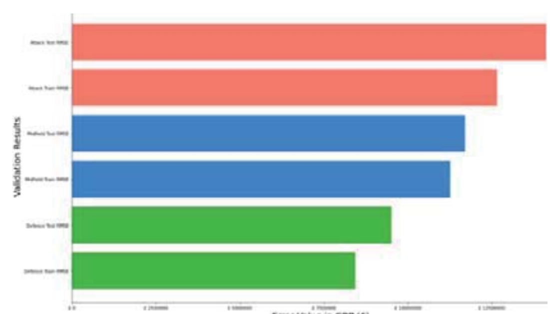

Fig.3.5.4 Gradient Boosting Regressor

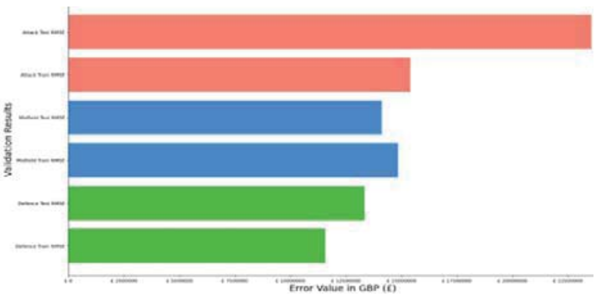

Fig. 3.5.5 Ada Boost Regressor

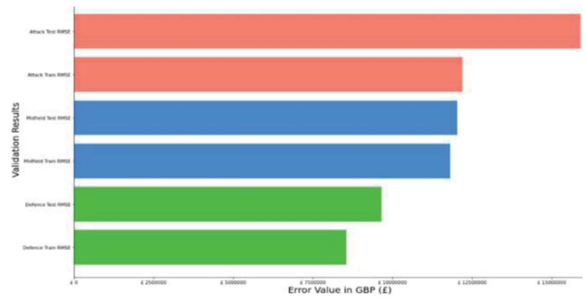

Fig.3.5.6 Support Vector Regression

Fig. 3.5 ML Models Graphs

4. Results and Discussions

Comparisons are performed between each model's results to identify which model worked best to predict player transfer values with the lowest RMSE for each zone. As seen clearly in Fig. (4.1.1), The Ridge (L2) Regression produced the lowest Root Mean Squared Error(£12m) in its predictions for attacking players. In the case of midfielders in Fig. (4.1.2), the Random Forest Regressor produced the lowest Root Mean Squared Error(£11.6m). And for Defenders, Fig. (4.1.3) shows the three ensemble models (Random Forest, Gradient Boost & Ada Boost) produced similar results but among them, the Random Forest Regressor was the model producing the lowest Root Mean Squared Error(£9.5m).

We finished our model with an error range of about £9–£12 million. The results of this experiment show that despite the fact that this model may be acceptable for larger teams looking to sign the top players, we believe that our models are inadequate for mid-tier and lower value players who are worth £20 million or less. As a result, we can infer that information on a player's performance on the field for these players alone might not be sufficient to predict the transfer worth of player with any degree of accuracy. For this to be accomplished, more data might be needed.

Fig. 4.1.1 Model Performance for Attackers

Fig. 4.1.2 Model Performance for Midfielders

Fig. 4.1.3 Model Performance for Defenders Position

Fig. 4.1 Model Performance for different Positions

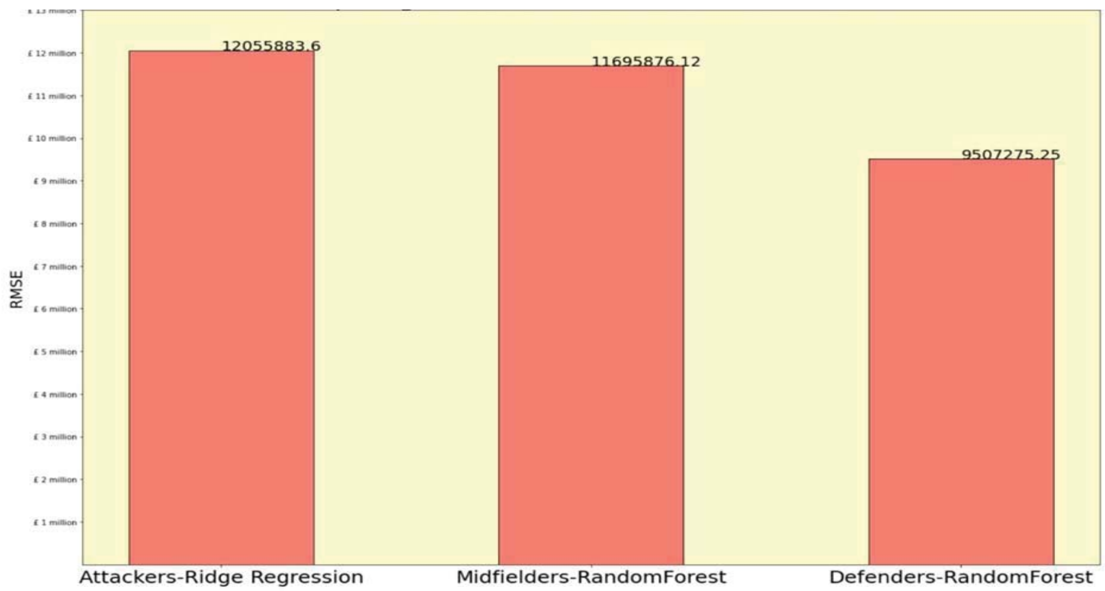

Fig. 4.2. Identifying which positions we predicted best

5. Conclusions

Through our findings, it is made evident that on-field performance, while extremely influential, is not the only predictor of a player's transfer value. That being the conclusion, it is important to have an understanding of a few restrictions that will need to be addressed given the limited scope for this project.

The first and foremost is the false presumption that the competition's top 5 leagues are all very contested. In addition, players on better teams typically have superior stats. One way of compensating for this could be through the creation of a "competitiveness score" based on leagues and opponents. Although time-consuming, this is doable because relevant data is obtainable.

Furthermore, in this project, players have been divided into positions as attackers, midfielders, and defenders. This has been done due to the sources the data has been obtained from still use these notations to compare players from different eras. Positions in modern football, however, are more nuanced. Players can further be classified according to different roles they play in a particular position (for example, attackers can be defined as centre-forward, winger, false-9 etc.). Comparisons made after the data is divided into these roles could yield more accurate results.

References

1. Carmichael, F., Forrest, D. & Simmons, R. (1999): The Labour Market in Association Football: Who gets Transferred and for how much? Bulletin of Economic Research, 51 (2), pp. 125-150
2. Carmichael, F. & Thomas, D. (1993): Bargaining in the transfer market: theory and evidence, Applied Economics, 25 (12), pp. 1467-1476
3. Muller, O., Simons, A., and Weinmann M., "Beyond crowd judgments: Data-driven estimation of market value in association football.", European Journal of Operational Research Volume 263, Issue 2, pp. 611-624 (2017).
4. Pujol, F., Garcia-del-Barrio, P. & Elizalde, J. (2007): ESIrg Report on Media Value in Football: The English Premiership Extends its Supremacy, Universidad de Navarra.
5. Frick, B. (2007): The Football Players' Labor Market: Empirical Evidence from the Major European Leagues, Scottish Journal of Political Economy, 54 (3)
6. Herm, S., Hans Markus Callsen-Bracker, HenningKreis, "When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community", Sport Management Review, Volume 17, Issue 4, pp. 484-492 (2014).
7. Hucaljuk, J, and Rakipović, A., "Predicting football scores using machine learning techniques", MIPRO 2011, Opatija, Croatia (2011).
8. Franck, E., Nüesch, S., "Talent and/or popularity: what does it take to be a superstar? "Economic Inquiry, 50 (1) pp. 202-216 (2012).
9. Tunaru, R., Clark, E. & Viney, H. (2005): An Option Pricing Framework for Valuation of Football Players, Review of Financial Economics, 14, pp. 281-295
10. Kesenne, S. Revenue sharing and absolute league quality; talent investment and talent allocation. Scott. J. Political

Econ. 2015, 62, 51–58.
11. Felipe, J.L.; Fernandez-Luna, A.; Burillo, P.; de la Riva, L.E.; Sanchez-Sanchez, J.; Garcia-Unanue, J. Money Talks: Team Variables and Player Positions that Most Influence the Market Value of Professional Male Footballers in Europe. Sustainability 2020, 12, 3709. https://doi.org/10.3390/su12093709
12. Erik van den Berg, "The Valuation of Human Capital in the Football Player Transfer Market: An investigation on transfer fees paid and received in the English Premier League", M.Sc. Thesis, Erasmus School of Economics, Erasmus Universiteit Rotterdam, July 2011.
13. Ian G. McHale, Benjamin Holmes, Estimating transfer fees of professional footballers using advanced performance metrics and machine learning, European Journal of Operational Research, Volume 306, Issue 1, 2023, Pages 389-399, ISSN 0377-2217, https://doi.org/10.1016/j.ejor.2022.06.033, https://www.sciencedirect.com/science/article/pii/S0377221722005082
14. Sindhwani, N.., Anand, R.., S., M.., Shukla, R.., Pratap Yadav, M.., & Yadav, V.. (2021). Performance Analysis of Deep Neural Networks Using Computer Vision. EAI Endorsed Transactions on Industrial Networks and Intelligent Systems, 8(29), e3. https://doi.org/10.4108/eai.13-10-2021.171318
15. Deal, B., Grove, A., 1965. General Relationship for the Thermal Oxidation of Silicon, Journal of Applied Physics 36, p. 3770.
16. Fachinger, J., 2006. Behaviour of HTR Fuel Elements in Aquatic Phases of Repository Host Rock Formations. Nuclear Engineering & Design 236, p. 54.