Comparative Analysis of Data Driven Techniques to Predict Transfer Prices of Football Players

Shubh Harde [#1], Vedant Bhawnani [*2], Prof. Shruti Savant [#3]

***Abstract*— Football clubs from all around the world are in constant business of buying and selling players each transfer window. These transactions of signing and selling players involve multi-million dollar deals. Therefore, it is essential for buyer clubs to estimate the cost of acquiring the services of a player they have set their eyes upon before they make the decision of spending millions of dollars for that particular player. This need has caught the attention of researchers, statisticians and enthusiasts of the sport, which has led to the development of several techniques and platforms which predict the how much the player would cost. Transfermarkt is one such platform which relies heavily on its community to decide market values of players. This assessment is subjective and results in inconsistencies. As a result, there is a proliferation in the number of data-driven techniques being developed to statistically predict price of players. In this paper, we review and compare such several data-driven techniques for predicting player prices in the footballing world.**

Keywords – Football, Transfer window, data-driven techniques, prediciting player prices, TransferMarkt,

1. Introduction

Football is known as the world's sport as it is the most popular sport in the world [1]. There are more than 200 professional football leagues all around the globe [2]. Among these leagues, some renowned leagues are England's Premier League, Spain's EA Sports La Liga, Italy's Serie A, Germany's Bundesliga and France's top flight, Ligue 1. All these leagues have 15-20 teams competing for the title each season. Put together, thousands of players play in the football leagues all over the world, each year. Players move around from team to team on a regular basis. When a player wants to make a move from one team to another, the two major parties involved are, the engaging team and the releasing team who often work to reach an agreement through an intermediary known as the player agent.

For any player, there are two types of fees, which are often used interchangeably, the transfer value and the market value. In reality, these are two distinct values. The market value is defined as the financial value of the player based on performances. The transfer value is defined as how much the player would cost the buying team. Transfer fees are heavily influenced by factors such as the player's image, the player's relationship with their current team, the shirt selling and fan pulling power of the player along with the length of the current contract and age of the player. In some cases, the transfer fees also include agent fees. The standing transfer fee record is €222 million paid by Paris Saint Germain to F.C. Barcelona for acquiring the services of Neymar Jr. in 2017.

Not over paying for any player is always a priority for a club trying to sign new players to strengthen the team. Over paying for an underperforming player could cause catastrophic harm to the club both, financially and sportingly. Hence, it is important for decision makers at a club to predict and evaluate the market value and transfer value of players.

This paper exhibits a thorough analysis of past work done for predicting the market value and transfer prices of football players. We first review the existing literature on the topic of transfer and market fees, describe all the data sources put to use and machine learning techniques used in the literature. Followed by putting forth the areas of improvement in the literature that we think can be worked upon by other researchers.

2. Literature Review

The prediction of football players' market value has been studied widely, with researchers employing different methods, datasets and algorithms to enhance prediction accuracy. This section encapsulates a comprehensive literature review focused on the review of literature delving into the prediction of market value of players through machine learning and deep learning algorithms. Some researchers have found out some parameters which play an important factor in predicting the market value of a player. Age is included as one of the important factors, with young players demanding more value due to their potential for growth. Similarly, players with high popularity tend to bring their fans with them, leading to an increase in their market value.[6] The position of a player is also thought to play an important role, with attackers usually getting more importance.[5]

Beginning with Carmichael & Thomas (1993), economists have used regression models to identify the determinants of transfer fees. Since then several researchers have followed a similar trend. Mustafa A. Al-Asadi et al.[6] used data from FIFA as true values and used different linear and non-linear models to predict the market price of a player. They employed 9 parameters, such as international reputation, weak foot column, etc.

Stanojevic and Gyarmati[4] presented their research on statistical measures, and obtained data from sports analyst company InStat, and transfermarkt (TM). The research aimed to estimate the market value of 12,858 players based on player performance metrics. The models were built on 45 predictors, and their results outperformed widely used transfermarkt.com market value estimates.

Muller et al.[7] presented a data-driven approach to overcome the limitations of crowd-sourcing. The researchers used the data from top 5 European leagues. They created a dataset using the attributes such as player characteristics - age, position, nationality. Muller et al.[7] took a unique approach by including data from Wikipedia, Facebook and Google metrics at that time. They employed a linear regression model, and results achieved were within the scope of crowdsourced estimates.

Dobson et al.[8] explored the effects of player metrics on transfer fees and discovered that transfer fees are volatile across segments even in a single competition. More recently, Depken II & Globan[9] use linear regression to identify that English clubs pay a premium in the transfer market, compared to clubs from other European countries.

Yigit et al.[10] presented an innovative approach to assessing the player values. They leveraged a wide range of player attributes, including on-field performance metrics, demographic data, and market factors, to predict player market values. The dataset comprised football players from major leagues. 5316 players from 11 major leagues across Europe and South America were considered. Data from the football manager simulation game was collected and merged with the transfer value from Transfermarkt. The most resultant values were in accordance with the current market values.

Behravan et al [11] took a distinctive approach to predict the market values of a player, by employing Particle Swarm Optimization. The data collected was from the FIFA 20 dataset, and the value of a player in the dataset was considered the true value. The players were divided into 4 clusters based on positions using an automatic clustering algorithm. According to the authors, the RMSE and MAE for their method are 2,819,286 and 711,029,413, respectively, while the results by Muller [7] were 5,793,474 and 3,241,733. These results indicate that their methods had a significant advantage over other methods [10]

Ian et. al[3] investigated the use of machine learning to estimate transfer fees, utilizing data from sofifa.com and transfermarkt.com. They trained both linear regression and XGBoost models on a range of performance metrics, including those from Instat and GIM performance ratings. Their findings indicated that the XGBoost model outperformed the linear regression model in predicting transfer fees. This research highlights the potential of machine learning to inform transfer decisions, addressing the question of "what is the expected fee of a player given their past performance?" The authors suggest further work to assess the "reasonableness" of transfer fees based on post-transfer performance, potentially leveraging the same data sources and machine learning techniques.

3. Data Sources

The quality of data plays a significant role in determining the real life performance of the algorithms. The model learns the existing data and outputs new responses, based on patterns found in the fed data. Hence, the data used needs to be not only accurate but also diverse. It needs to take into account a variety of possible factors that may be helpful in determining the output. In this section, we will discuss the different datasets/data sources used over the years to predict the transfer prices or the market price of a football player.

Several papers make use of performance metrics and event data collected from professional leagues, which was collected by scraping web pages. Other papers follow a more statistical approach for their analysis. Behravan and Razavi[11] looked to address the drawbacks of transfermarkt market value by using data from FIFA 20, by EA Sports. A few other authors followed a similar approach by using FIFA game data, such as Mustafa et al[6]., Vinscent Stevel et al., V. B. Jishnu et al.[12] Each of these authors use data from FIFA game, and consider those values to be the real values. Furthermore, some authors also specify the reason for using this dataset being a lack of real world statistical data for niche players.

A lot of research done in this field relies heavily on the transfer data from TransferMarkt(TM), and statistical data from fbref.com. TransferMarkt has been widely used due to the crowd-sourced player valurations. Members of the website offer their valuations of players and a panel of experts calculates a weighted average of the values to arrive at a single transfer value for each paper. The panel of experts calculate the weights based on judging how accurately each member has valued players historically. fbref.com provides a comprehensive study of football players, from basic information such as age, height and nationality, to more statistical information such as xG.

Another online source used for data analysis is sofifo.com. Player ratings have most commonly been taken from sofifo.com website. Members of the website provide ratings of players in many attributes(passing, shooting etc.), with editors reviewing these before presenting single values for each player. Yigit, Samak & Kayak[10], and Behravan & Razavi[11], both use crowd-sourced sofifa player ratings.

Other data sources include InStat, used by McHale & Holmes[3].

A study by Stanojevic and Gyarmati[4] relied on player tracking data and match event data, gathered from various leagues, to assess the technical aspects of players' performances. This data-driven approach allowed the researchers to estimate player values more accurately by focusing on specific in-game actions.

Other papers, such as that by Dobson and Gerrard[8], use historical transfer fee data from English soccer, which provides a comprehensive view of how player values have evolved over time in one of the sport’s most commercially significant leagues.

Apart from the data sources, another interesting aspect is the diversity of data used. While Depken II & Globan[9] use data from Europe’s top five leagues. Qing et al.[5] uses the UEFA championship performance data. Stanojevic and Gyarmati[4] also relied on player tracking and match event data, which provided a new way of thinking about the data that can be used. McHale & Holmes[3] use data from nine seasons starting with the 2011/12 season, till the 2019/20 season.

By leveraging the data obtained from these sources, ranging from professional metrics to crowd opinions to game statistics, they provide a holistic view of predicting transfer fees of a player. The growing amount of data makes it easy to envision the similar projects will be able to use the data for innovative and elegant purposes like transfer value prediction and other analysis.[11]

4. Algorithms Used

Linear Regression: Linear Regression assumes a linear relationship between the input variables (the features) and the output variable (the target). McHale & Holmes[3] have used it to model the relationship between on-field performance metrics and their transfer fees Where y is the transfer fee, x1, x2,.....,xn are the independent variables and E is the error term. In McHale & Holmes[3] Linear Regression provides a baseline model to learn how individual features of players (like age and goals per game) are related to transfer fees. However this technique struggles to capture complex non-linear relationships. Al Asadi and Sakir Tasdemir[6] also use Linear Regression to provide a rudimentary modelling of the relationship between players' attributes and market values. Only the players' potential is used as an independent variable in their study for this model. Linear Regression achieves a low accuracy with RMSE of 5.46 and R-Squared of 0.43.

Elastic Net Generation: The ENG regression mixes Lasso and Ridge regressions to handle the multicollinearity efficiently and perform variable selection. Elastic Net Generation is used in McHale & Holmes [3] for selecting only the most important features and managing overfitting. This is useful when there are many correlated variables.

XGBoost (Extreme Gradient Boosting): XGBoost is an ensemble learning technique which builds a series of decision trees, where each tree is aimed at correcting the errors of the previous one. In this technique, each tree learns from the mistakes of the previous trees, reducing the residual errors step by step. L1 and L2 regularisation both are used to prevent overfitting. XGBoost identifies the most important features that impact the target variable which makes the model more interpretable. McHale & Holmes [3] use XGBoost because it is very capable at capturing complex, non-linear relationships between player performance and transfer fees. XGBoost achieves the best performance compared to other methods. Yigit Samak and Kaya[10] also used XGBoost to improve prediction accuracy by adding trees that correct errors of previous trees, while controlling for overfitting. XGBoost achieved the lowest MSE of 0.170, making it the best performing model in the study. Jishnu et al.[12] used XGBoost to predict transfer values of players using the FIFA 22 video game dataset and other advanced player performance metrics. The model performed well across all categories of players with an R-Squared value of 0.805, 0.805, 0.81 and 0.856 for goalkeepers, defenders, midfielders and attackers respectively.

XGBDART (Dropout Additive Regression Trees): XGBDART extends XGBoost by adding a "dropout" technique which is inspired by neural networks, to randomly drop trees during the training phase. This prevents certain trees from dominating the model, resulting in better generalisation and less overfitting. McHale & Holmes [3] apply XGBDART as a way to regularise the boosted trees, improving generalization on unseen data.

Mixed Effects Linear Model: Mixed Effects Linear Model is useful when the data is hierarchical or grouped data because it allows for random effects to learn group-level variability. In McHale & Holmes[3] clubs are considered to be groups and the influence of transfer fees is treated as random effects. The authors use this model to model the varying financial behaviour of both, the buying and selling clubs while estimating transfer fees of the players. By doing this, the variability in the club's strategies is better captured.

Random Forests: Random forests construct multiple decision trees during the training phase and output the mean prediction for regression tasks. Here, each decision tree is trained on a random subset of data, which helps in reducing overfitting. The final prediction is an average of all trees' output. Stanojevic and Gyarmarti[4] use performance metrics like goals, assists, tackles and passes to predict players' market values. Accuracy and performance of random forests in predicting values is compared to other algorithms in the paper. Al Asadi and Sakir Tasdemir[6] have also incorporated random forests in their study to obtain a final prediction of players' market values. Random Forests perform better than the other algorithms in the study with an RMSE of 1,64 and R-Squared of 0.95. Random Forests were also used by Jishnu[12] et al. and had an R-Squared 0.783 for goalkeepers, 0.78 for defenders, 0.751 for midfielders and 0.783 for attackers. However, in Jishnu et al[12]. Random Forests fall short of other algorithms used.

Gradient Boosting Trees: GBT builds models in a sequence by minimising the residual errors of previous models. GBT is a more powerful and flexible model compared to Random Forest as it corrects the mistakes of the previous trees. GBT is particularly useful when dealing with complex and non-linear relationships of the data. Stanojevic and Gyarmarti[4] use GBT to predict market values of players taking into account a wide array of player statistics and correcting them, for inconsistencies. GBT is used as the main model for predicting the market value, which is attested against the actual values from Transfermarkt. In Jishnu et al,[12] GBT performed the best in the attackers category by achieving a R-Squared of 0.86, and had a strong performance in other categories as well with R-squared values of 0.813 for goalkeepers and 0.788 for defenders.

Coefficient of Variation (CV): The CV is the ratio of the standard deviation to the mean of a player’s performance metrics. This provides a normalized measure of variability which allows comparisons across different players and performance metrics. In Yi, Qing et al.[5] CV is applied on player statistics to evaluate match-to-match performances for players across multiple positions and contexts. Higher the CV, more the variability in performance, while lower variability suggests more consistent performance.

Magnitude Based Inference: MBI is a statistical approach that is used to estimate the possibility of observed features being significant according to predefined thresholds. In Yi, Qing et al.[5] MBI is used the measure the difference between players’ performance spanned over several situational contexts.

Effect Size with Confidence Intervals: Effect Size is computed to understand the magnitude of difference in players’ in match performances across multiple categories. Effect Sizes calculate the amount of difference, independent of sample size. In Yi, Qing et al.[5] the effect sizes are calculated to compare the variability of player performance metrics over several different situations, such as home vs. away games.

Multiple Linear Regression: Multiple Linear Regression is also used by Al Asadi and Sakir Tasdemir[6] to model more relationships between the market values and player attributes. Several independent variables are used such as age, height, potential, international reputation, weak foot, team position, shooting, passing and dribbling. MLR shows an improvement over linear regression with an RMSE of 4.66 and R-Squared of 0.56. MLR is also used by Yigit Samak and Kaya[10] to predict market value of players. They use it as a baseline model which produced a mean squared error of 0.768 when validated using cross-validation.

Decision Trees: Al Asadi and Sakir Tasdemir[6] used to recursively split the data into small subsets of player attributes. To minimise the error in predicting the market values of the players, the tree keeps branching based on the most significant features which helps in handling non-linearity. Decision trees significantly improved upon the performance of linear and multiple linear regression by achieving a RMSE of 2.71 and R-Squared of 0.87.

Multilevel Regression Analysis: Muller Simons and Weinmann[7] employed this algorithm to make an estimation of market values of football players by considering multiple features including player characteristics, performance metrics, and popularity data. This model was selected by the authors because the data is hierarchical (players nested in teams and teams nested in leagues) as well as longitudinal (spanned across multiple seasons). This model achieved a +3.4% relative difference of RMSE between crowd estimates and model estimates and a +3.6% difference of MAE between crowd estimates and model estimates

Ridge Regression: Used in Yigit Samak and Kaya[10], Ridge regression is an improvement upon linear regression which regularises the coefficients of less significant features, thus preventing overfitting. The goal of ridge regression is to shrink the coefficients of the insignificant variables towards zero while still keeping the coefficients of the significant features intact. The MSE of the ridge regression model was 0.608.

Lasso Regression: Yigit Samak and Kaya[10] also use Lasso regression goes one step further than ridge regression and sets the coefficients of insignificant features to exactly zero. The MSE of Lasso regression was 0.590

APSO-Clustering (Automatic Particle Swarm Optimization Clustering): APSO-Clustering is an automatic clustering method which consists of two steps. The first step involves breaking down large and complex datasets into optimal clusters and the second step identifies the correct positions of the clusters obtained at the end of the first step. Iman Behravan and Mohammad Razavi[11] used APSO-Clustering’s first phase to cluster players based on their positions. This helped ensure that the models applied later are specific and relevant to each position.

PSO-SVR (Particle Swarm Optimization - Support Vector Regression): PSO-SVR is a hybrid machine learning model which is made by combining Particle Swarm Optimization and Support Vector Regression. In this hybrid model, PSO performs two tasks, feature detection and parameter tuning of SVR. PSO iterated through the search space to identify the most significant subset of features and optimises SVR’s kernel parameters. In Iman Behravan and Mohammad Razavi’s[11] PSO-SVR was applied to each cluster to estimate the market values of players. PSO-SVR achieved a accuracy of 74%, which outperformed other algorithms like GWO (Grey Wolf Optimizer) (70% accuracy), IPO (Inclined Planes System Optimization) (67% accuracy) and WOA (Whale Optimization Algorithm) (64% accuracy).

5. Conclusion and Future Work

In this paper, we have provided an in-depth study and analysis of the existing literature from the year 1993 to 2023 of predicting football players’ market and transfer values. Our review compasses of 10 research papers which made use of 15 different machine learning techniques to estimate the market and transfer values. Out of all the models used in the vast literature, XGBoost has proven to be the best performing model across multiple papers as it achieved a low MSE of 0.170 in Yigit Samak and Kaya [10], and an R-Squared of 0.805 in Jishnu et al [12]. On the other end of the spectrum, linear regression came out to be the worst-performing model in the referred works. Linear regression was often used as a baseline model in the literature because it fails to capture non-linear relationships in the datasets.

In future research, we plan to develop a comprehensive Transfer Value Index (TVI) which will serve as a single metric required to estimate the transfer value of football players. This index will cover various dimensions that are involved in deciding the transfer values of players. The index will combine advanced in-match statistics – such as xG, PsXG, xGA, passing accuracy, and tackling accuracy – with off-field factors that have a significant influence on transfer fees. The player’s reputation with both the fans and the club, including media presence and fan pulling power will also be used to reflect the player’s brand value. Finally, we will also take into account the economic as well as the sporting position of buying and selling clubs. Therefore, by combining these variables into a single index we aim to offer a holistic, data-driven and accurate estimation of player transfer fees which would serve as a valuable tool for clubs, agents and fans in future player transfers.

6. REFERENCES

[1] Eleni Veroutsos. (2023). The Most Popular Sports In The World- worldatlas.com.URL: https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html

[2] Top 10 Leagues in the World URL: https://jobsinfootball.com/blog/best-soccer-leagues-in-the-world/

[3] McHale, Ian G. & Holmes, Benjamin, 2023. "Estimating transfer fees of professional footballers using advanced performance metrics and machine learning," European Journal of Operational Research, Elsevier, vol. 306(1), pages 389-399.

[4] R. Stanojevic and L. Gyarmati, "Towards Data-Driven Football Player Assessment," 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 2016, pp. 167-172, 10.1109/ICDMW.2016.0031.

[5] Yi, Qing et al. “Situational and Positional Effects on the Technical Variation of Players in the UEFA Champions League.” Frontiers in psychology vol. 11 1201. 19 Jun. 2020, 10.3389/fpsyg.2020.01201

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

[7] Müller O, Simons A, Weinmann M. Beyond crowd judgments: data-driven estimation of market value in association football. Eur J Oper Res. 2017;263(2):611–24. 10.1016/j. Ejor.2017.05.005.

[8] Dobson S, Gerrard B. The determination of player transfer fees in English professional soccer. J Sport Manag. 1999;13(4):259–79. 10.1123/jsm.13.4.259.

[9] Depken II, C. A., & Globan, T. (2021). Football transfer fee premiums and Europe’s big five. Southern Economic Journal, 87(3), 889–908. 10.1002/soej. 12471.

[10] Yiğit, A.T., Samak, B., Kaya, T. (2020). Football Player Value Assessment Using Machine Learning Techniques. In: Kahraman, C., Cebi, S., Cevik Onar, S., Oztaysi, B., Tolga, A., Sari, I. (eds) Intelligent and Fuzzy Techniques in Big Data Analytics and Decision Making. INFUS 2019. Advances in Intelligent Systems and Computing, vol 1029. Springer, Cham. 10.1007/978-3-030-23756-1_36

[11] Behravan I, Razavi SM. A novel machine learning method for estimating football players’ value in the transfer market. Soft Comput. 2021;25(3):2499–511. 10.1007/ s00500-020-05319-3.

[12] V. B. Jishnu, P. V. H. Narayanan, S. Aanand and P. T. Joy, "Football Player Transfer Value Prediction Using Advanced Statistics and FIFA 22 Data," 2022 IEEE 19th India Council International Conference (INDICON), Kochi, India, 2022, pp. 1-6, 10.1109/INDICON56171.2022.10040117.