Player Replacement Finder

Abstract

The transfer markets in football are a dynamic and pivotal period which entails a state of volatility where teams and clubs strategize and aim to strengthen their teams. This is more often than not, preceded by in depth analysis of loads of players to find the perfect replacement for the outgoing player. This analysis in recent times includes labour intensive tasks of manual video reviews. Our work introduces the Football Player Replacement Finder, a novel approach to reduce the complexity and time required for scouting and acquiring impactful talents by using advanced machine learning models and automated data scraping pipelines. Our system employs supervised models for gauging the performance and price of football players along with clustering techniques for player profiling, enabling stat-by-stat comparison of players. By integrating advanced metrics along with appealing visualisations, our system empowers decision-makers to streamline their scouting process and uncover valuable talents ef ectively.

1. Introduction

The transfer window months of january and july each year, are one of the most critical time period for any club looking to strengthen the team. In this period, the clubs focus heavily on identifying potential signings, while keeping a close eye on the departing players. Each club has to identify suitable replacements while ensuring that they do not overpay for a player as adhering to the budget is a key requirement for any club. Scouting players and ensuring that the scouted players are within the budget, is a monumental task that requires a deep analysis of a wide range of factors, such as player statistics, market conditions such as inflation and volatility, the team effect on team dynamic of incorporating a new player, and any ego clashes that may arise in the dressing room among players. Traditionally, this process includes extremely laborious scouting methods, involving going through extensive video footage, gathering subjective reports and insights, and manually comparing hard statistics of players. These methods have always been time consuming and with the exponential increase in the amount of data and advanced metrics, it is only getting increasingly difficult to keep up with the demands of modern football. The recent advancements in football analytics have introduced multiple data-approaches to player evaluation. Platforms such as FBref and Sofifa offer vast amounts of data and player statistics and attributes, which have become an important tool for the recruiters. Despite the availability of these insights, the process of finding replacements remains labour-intensive. Traditional methods of scouting do not generally incorporate data driven approaches while handling larger and more complex datasets in the modern era of players scouting. We have designed an advanced system of player recruitment and replacement process by automating data collection, analysis and comparison. Our system uses machine learning models, including supervised learning and clustering algorithms to predict player performance and identify statistically similar players. By making use of similarity metrics such as Euclidean distance, cosine similarity or the Pearson Correlation, the system enhances the accuracy of player comparisons, ensuring a more relatable match between the outgoing player and the potential replacement of the outgoing player, The integration of automated data scraping pipelines allows us faster and more effective comparison of players based on on-field metrics such as expected goals, expected assists, goals, assists, passing accuracy, progressive carries and other related statistics. These metrics, combined with similarity
measures, creates a comprehensive view of player performance, enabling deeper and more accurate

analysis, ensuring that the players are suitable not only for the immediate requirements of the club, but also with the tactical approach of the manager and the coaching staff along with the long-term goals of the club.

By reducing the reliance on manual processes, we eliminate the risks of human errors, and biases that are brought into the system by the recruiter's prejudices. *Player Replacement Finder* aims to enable clubs in easier and faster identification of potential players with greater accuracy. The proposed system not only enhances the precision of player comparisons models, but also offers intuitive visualisation approaches that helps clubs take faster decisions in choosing the right candidate and saving them from over spending unnecessarily. This approach has the potential to optimise transfer strategies, which indirectly also affects the team performance.

The remainder of this paper is organised as follows: Section 2 reviews the current literature on football player replacement and data analytics already accomplished in this domain. Section 3 outlines the methodology used to develop the *Player Replacement Finder* system, including the data sources and machine learning models employed. Section 4 presents the results and validations of the system, and section 5 concludes with discussion of the system's impact on modern football transfers and potential future development.

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 predicting the market value of players through machine learning and deep learning algorithms. Few 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 a player plays is also thought to be important, with attackers usually having higher market values..[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, with many researchers following regression-based models to forecast the player transfer fees. Mustafa A. Al-Asadi et al(2022).[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, following a more traditional approach, aimed to estimate the market value of 12,858 players based on player performance metrics. Clustering techniques were deployed to analyze player performance data, and the model was built on 45 predictors, with their results outperforming widely used transfermarkt.com market value estimates, paving way for more precise predictions with the help of data analytics and finer data.

Muller et al.[6] 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. 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.[7] 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[8] 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.[9] 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.[10]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[6] 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. Methodology

The proposed methodology is divided into three main sections -

1) Data pipeline

2) Classification and Similarity Search

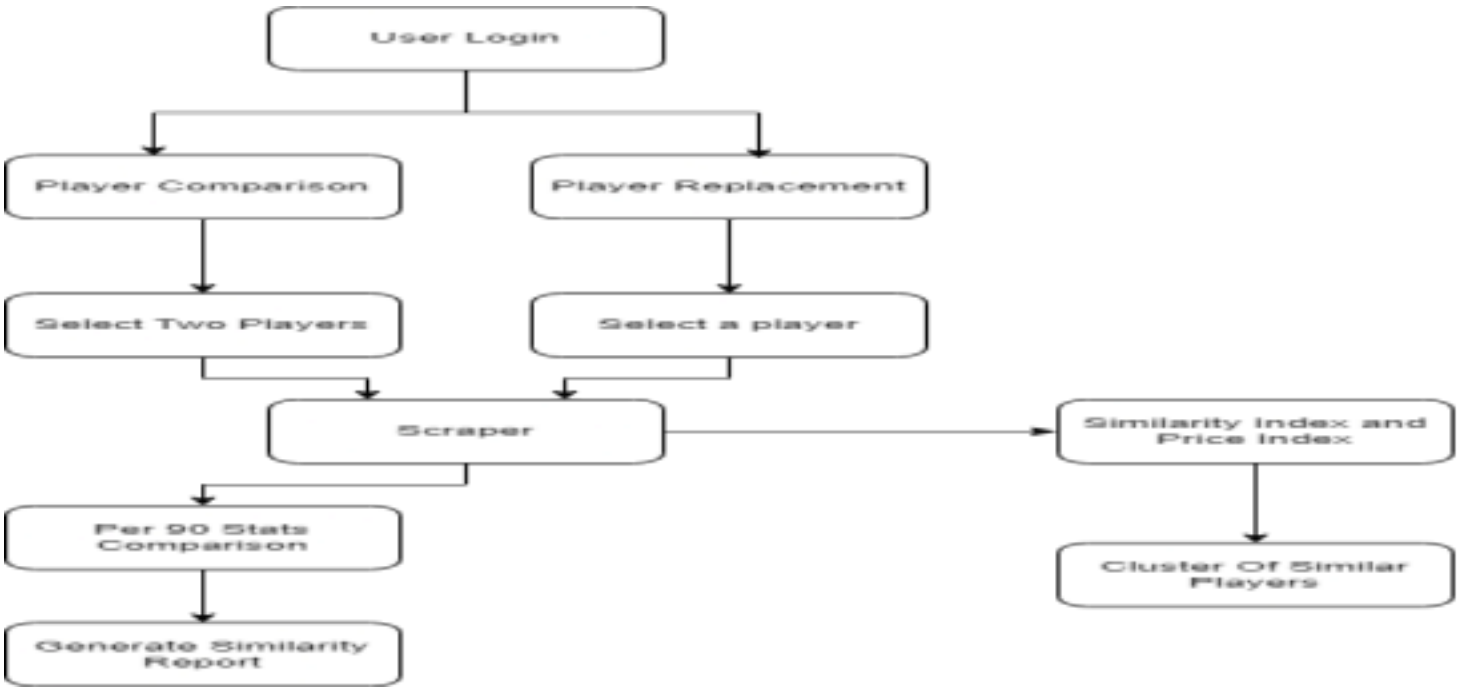

Figure 1: Methodology of proposed system

3.1) Data Pipeline

The data pipeline serves as the foundation of the system, ensuring that data is systematically collected, cleaned, and stored for efficient use in subsequent analysis. This process automates the process of extraction and preparation of data for the tasks of classification, similarity analysis, and regression modelling.

3.2) Data Collection

The first step in the pipeline involves collection of data from external sources.The data was collected primarily from FBRef and TransferMarkt, which provide a comprehensive view into the football statistics for a player, including, but not limited to, personal information such as height, birth country, and team-related information (e.g., last played club, current club), positional statistics such as xG for forwards, and tackles won for defenders.TransferMarkt provided positional data along with crowd-sourced market prices for a player, and their current injury status.

TransferMarkt:

FbRef:

3.3) Data Preprocessing

Ensuring data integrity and ensuring clean data is essential for machine learning models, and for improving the accuracy for these models. To ensure that the data we use is cleaned, the following steps were implemented:

- The data from fbref and TransferMarkt were merged based on the player last name, and age column. The age column was chosen because it was present in both, scraped tm(TransferMarkt) data, and fbref data. This ensured that the merging was correct, and players with the same name were handled correctly.
- Given the presence of missing values for position specific stats, for example a goalkeeper did not have xG, and fbref did not have data on some new players, any row with more than 85% missing values was dropped. This ensured that the dataset remained robust. For the remainder of the dataset, the numerical values were imputed using the median where the missingness was minimal, else the mean of the column was used. Minimal feature engineering was performed at this stage, such as calculating the remaining contract length, and the experience of a player.
- Winsorization was performed on age and market value to handle outliers in these columns. It is a statistical technique used to limit extreme values in data to reduce the impact of outliers, involving adjusting outlier values to a specified percentile, thereby making the data more robust and reliable for analysis. It was noticed that winsorization impacted the final result, and normalizing the age for too long or too old players helped standardize the data. Finally IQR was applied to handle any outliers that may exist in the remaining columns.

1. Feature Engineering

   Based on all positions, positions were merged into 4 classes - forward, midfielder, defender and goalkeeper. Based on each specific position, specific attributes were created, from existing data. For example, ‘Defensive Actions’ was created for defenders, using the tackles, interceptions, blocks, clearances and recoveries, and normalized for per 90. Similarly about 10 statistics were created for each of the first three positions - forwards, defenders and midfielders. 5 statistics were created for goalkeepers that best defines the goalkeeping stats, based on the recent developments in the style of

play of goalkeepers.

3.4) Modelling

A suite of supervised models are applied to predict the market value of a player - driven by the studies presented by Ian et al., and Depken II et al. - such as Random Forest, Gradient Descent and XGBoost regressor. Linear regression was used as a baseline model, along with training an XGBoost on the cleaned dataset without any feature engineering to understand the importance of the created features. Moreover, stacking, bagging and boosting ensemble methods were also employed, to test if ensembling could help increase the R2 score. Stacking multiple regression models, along with PCA based on graphical methods and Kaiser's rule, helped increase the accuracy of the model, reaching a State-of-the-Art R2 score. This collaborative approach helps not just in precise market price prediction but also finds practical applications in scouting and transfer decisions by revealing underlying patterns with respect to player similarity and player performance.

IV. Results

V. Conclusion and Future Scope

The proposed system, Player Replacement Finder, provides a one stop solution to football's one of the most important stages, player scouting. Identifying players that suit the team's dynamic is a challenge, and this system aims to streamline that. By integrating pipelines, supervised learning, unsupervised learning for similarity indexing, and LLMS for report generation, it automates the labour intensive tasks, and saves time of the recruiters that could be better spent elsewhere. Player replacement finder utilises traditional game metrics, along with data-driven metrics, to simplify the decision-making process and help the coaches find players that better synchronise with the team. The system promises a significant leap in football analytics, offering clubs with a powerful tool at their disposal, saving the club vital time and money, while providing assurance about making the right choice.

VI. References

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

[2] Top 10 Leagues in the World - jobsinfootball.com [online]. Available at: 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, doi: 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, doi: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, doi: 10.1109/ACCESS.2022.3154767. keywords: {Sports;Games;Biological system modeling;Random forests;Measurement;Data models;Analytical models;Player value prediction;regression;machine learning;football analytics;FIFA video game data},

[6] 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. https://doi.org/10.1016/j. Ejor.2017.05.005.

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

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

[9] 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. https://doi.org/10.1007/978-3-030-23756-1_36

[10] 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. https://doi.org/10.1007/s00500-020-05319-3.

[11] 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, doi: 10.1109/INDICON56171.2022.10040117.