Techno-Science

Scientific Journal of Mehmet Akif Ersoy University

www.dergipark.gov.tr/sjmakeu

Review Article

hosted by Turkish JournalPark ACADEMIC

A SURVEY ON FOOTBALL PLAYER PERFORMANCE AND VALUE ESTIMATION USING MACHINE LEARNING TECHNIQUES

Vehbi Hakan SAYAN[1], Emrah HANÇER[2*],

[1] *Management Information Systems, Süleyman Demirel University, Turkey*
[2] *Department of Software Engineering, Burdur Mehmet Akif Ersoy University, Turkey*

ARTICLE INFO

Article History

Received	:	07/12/2022
Revised	:	30/12/2022
Accepted	:	30/12/2022
Available online	:	31/12/2022

Keywords
Football Video Games, FIFA, Football Manager, Machine Learning

ABSTRACT

The popularity of games like FIFA and Football Manager has attracted millions of players. Data plays an increasingly prominent role in these games. In other words, these games contain data from all soccer players worldwide, which can be used to simulate soccer games. Therefore, a lot of experts and a lot of tools are involved in producing football data. It is not only the production of this data that has attracted researchers, but also the analysis of it since useful and consistent information can be extracted from it to design real-world football applications. In this study, we introduce a survey of studies which focus on the prediction of player value/performance using machine learning techniques. As far as we know, there is no survey in the literature that specifically addresses this topic.

1. INTRODUCTION

As a sport branch, football spread across the world during the 20th century and became popular across the globe. Football became a major entertainment industry in the 21st century. In addition to sponsorship deals and TV deals, this expansion affected the market value of football players as well. A player's market value can be considered an estimate of how much a team would be willing to pay to get him to sign a contract, regardless of the actual deal [1]. The market value of a player is resulting from various factors like talent, popularity, skill, playing style, efficiency, etc. [2]. Therefore, players' values are still highly judgmental in the present-day football world. It is common for even the most richly resourced clubs in the world to transfer players for fees they regret later on. At the same time, a player who attracts high transfer fees for his small club can become the club's savior. For football clubs to be successful, both financially and competitively, transfer decisions are becoming increasingly crucial in this competitive environment. A machine learning algorithm may assist coaches and clubs in deciding which players to transfer. However, due to the difficulty of gathering detailed information about players, there has been limited research in football analytics using machine learning techniques.

Transfermarkt.com is the leading online source that uses crowd estimation to determine the value of players on the transfer market. A number of football-related data can be found on the site, such as results, scores, transfer rumors, football news and estimates of market value for most professional football leagues [4]. The method used in Transfermarkt.com involves members of the website estimating the players' values, which are determined by special members, known as mentors, based on the estimates of the other members. However, this method has following limitations. Firstly, inexperienced members can lead to inaccurate estimations. In addition, it is generally appropriate for well-known players. Thirdly, the value of players is not frequently updated since the system requires feedback from mentors. As an alternative to Transfermarkt.com, video games like FIFA and Football Manager (FM) are also considered as sources of data for football analytics. The use of video games as an alternative source of data has been gaining traction since 2014. It is an asset to have so much online data about football. However, it needs to be filtered and analyzed properly in order to make predictions about player value. Moreover, the process is not always straightforward. Additionally, incidents not captured in the data may influence a player's performance; for example, poor performances by opponents may cause a player to be rated higher than he deserves.

* Corresponding Author: ehancer@mehmetakif.edu.tr

To cite this article: SAYAN V.H., HANCER E., (2022). A Survey on Football Player Performance and Value Estimation Using Machine Learning Techniques, Techno-Science, vol. 5 no. 2- p. 57-62

e-ISSN 2651-3722

Various studies have been conducted with football data in the literature. The following are some areas of study: football player value estimation, match result prediction, predictions made using football video data, crowdsourcing for player value estimation, social networking analysis for player value prediction, etc. In this survey, we consider studies concerning player performance and value estimation with data based on FIFA video games. To the best of our knowledge, there is no survey in the literature that specifically addresses this topic.

The remainder of the paper is as follows. Section 2 describes the overall view of FIFA datasets. Section 3 presents the corresponding studies with discussions. Section 4 summarizes the corresponding studies.

2. STUDIES ON PLAYER PERFORMANCE AND VALUE ESTIMATION

FIFA is a worldwide video game developed by Electronic Arts (EA). In order to keep the data up-to-date, EA employs 30 data producers and 400 data contributors. Furthermore, the database is reviewed or modified by 6000 SOFIFA talent scouts and reviewers. The FIFA Soccer Games provide a comprehensive and coherent scout of players from all over the world. The game includes all the major leagues in the world with over 17000 players, 30 leagues, and 700 clubs. Simulated FIFA players are categorized into 14 positions and rated on 29 different skills, with each skill being evaluated on a 0-to-100 scale. As well as attributes such as mental abilities, physical abilities, field position, and other characteristics (cardinal variables) such as nationality, flag, etc, the data also includes monetary values, wages, and release clauses of football players. Some studies conducted on FIFA-based data are described as follows.

Lee et al. [5] examined the key factors affecting the transfer fees of top soccer players around the world based on FIFA data analysis. An improved LightGBM model was proposed by optimizing its hyper parameters using a Tree-structured Parzen Estimator (TPE) algorithm. A Shapley Additive Explanations (SHAP) algorithm was used to identify prominent features. Various baseline regression models (including linear regression, lasso, elastic net, and kernel ridge regression) and gradient boosting models without hyper parameter optimization were compared against the proposed method. The optimized LightGBM model outperformed the others.

Table 1. Summary of Articles

Reference	Research Purpose	Data Source	Features	Modelling Technique	Metrics
Lee et al. (2022)	Football player value estimation using bayesian ensemble approach	FIFA, WhoScored. com	124 features	Linear Regression, Lasso Regression, Kernel Rigde Regression, Elastic Net, Gradient Boosting Decision Tress, LightGBM	MAE ,RMSE
Chavan and Dondio (2019)	Applying machine learning techniques for football player selection	FIFA	43 attributes, 17589 players	Naive Bayes, Decision Tree, Support Vector Machines, XGBoost, K Nearest Neighbours	Accuracy, Precision, Recall, F1
Al Asadi and Taşdemir (2021)	Machine learning with balancing and feature selection techniques	FIFA	Base model 29 features, decreased to 10	Random Forest Classifier	Accuracy, Precision, Recall
Al Asadi and Taşdemir (2021)	Football player value estimation using machine learning techniques	FIFA	7 features	Linear Regression, Multiple Linear Regression, Decision Trees, Random Forests	MAE, RMSE, R^2
Yaldo and Shamir (2017)	Football Player wage calculation, and analysis according to different leagues, different positions by using machine learning techniques	FIFA	40 features	Additive Regression, Decision Table, Nearest Neighbor with a weighted condition, K*, Locally Weighted Learning with Naïve Bayes and Linear regression classifiers, Random Committee, Random Trees and Random Subspace	MAE, R
Pariath et al. (2018)	Football player performance prediction using machine learning	FIFA	36 features, 21280 players (4 dataset for 4 player positions created, 6000-9000 players)	Linear Regression, PCA	MAE, RMSE, R^2, Median Absolute Error
Yigit et al. (2020)	Football player value estimation using machine learning techniques	Football Manager, www.trans fermarkt.c om	55 features, 5316 players	Validation Set, Cross Validation, Ridge Regression, Lasso Regression, Principal Component Regression and Partial Least Squares, Decision Tree, Random Forests and Extreme Gradient Boosting, Ensemble Model (0,7*Decision Tree + 0,3*Regression)	MSE

Table 2. Summary of Articles

Reference	Research Purpose	Data Source	Features	Modelling Technique	Metrics
Apostolou and Tjortjis (2019)	Player position, number of goals for season and number of shots per match calculation	FIFA	-	Random Forest, Multi Layer Perceptron Classifier, Linear Support Vector Classifier	Confusion Matrix, Number of goals, Number of shots
Rajesh et al. (2020)	Team formation and different statistics by machine learning	FIFA	36 features	Naïve Bayes, Random Forest, Decision Tree, Support Vector Classification, one proposed prediction algorithm. K-Means	Accuracy, F1, Jaccard Similarity
He et al. (2015)	Football player's performance and market value	www.transfermarkt.com, WhoScored.com, European Football Database and Guardian voting system	100 features for value 84 for voting	Lasso Regression	MSE, Developed Metric
Stanojevic and Gyarmati (2016)	Player market estimation and comparison with real values using machine learning methods	www.transfermarkt.com, InStat	45 features, 12858 players	Random Forests, Gradient Boosting Trees Regression and generalized linear models	RMSE, R, Median Error
Rao and Shrivastava (2017)	Team formation with machine learning algorithms	FIFA	30 features, 17800 players	Artificial Neural Network, Multinomial Logistic Regression, Random Forests	Accuracy, Precision, Recall, F1
Uzochukwu et al. (2015)	Football player selection with machine learning	www.pesstatsdatabase.com	30 features	Neural Networks	Metrics developed and average status calculated
Behravan and Razavi (2020)	Football player estimation with particle swarm optimization and support vector regression	FIFA	55 features, in different numbers after APSO	Support Vector Regression	MSE, RMSE, MAE, MAPE, R^2
Cotta et. al (2016)	Clustering for soccer analytics	FIFA	19 features	Linear Regression, PCA, K-Means	-

A methodology was proposed by Chavan and Dondio [6] to locate the most appropriate player within a short period of time for a player who moved to another club. In the preprocessing stage, data cleaning and feature engineering processes were applied to prepare the data for further process. In the second stage, data transformation (e.g. scaling and binary to numeric attribute conversation) was performed to transform the data into an appropriate format for a classifier. In the final stage, a variety of classification methods, including support vector machine (SVM), linear discriminant analysis (LDA) and k-nearest neighbors (KNN) were applied using the 10-fold cross-validation technique to determine the most suitable player instead of a player moving to another club. According to the results, SVM achieved the best scores.

Al-Asadi and Tasdemir [7] introduced a methodology to characterize football players in terms of nine field positions. In the first stage, data resampling techniques were applied to address class-imbalance problems in the data. In the second stage, feature selection techniques were conducted to deal with high-dimensionality. Finally, a classification process was conducted using a variety of classifiers. Although feature selection and resampling techniques have a significant impact on the classification performance, the methodology requires improvements to achieve high prediction performance.

In another study, Al-Asadi and Tasdemir [8] proposed a methodology to estimate the market value of players. In the first stage, redundant features were eliminated, and missing values were filled. Moreover, the positions of football players were categorized as three categories, namely forwarder, midfielder and defender. In the second stage, Pearson correlation

analysis was conducted to identify the relationships between features. Finally, linear regression process was conducted by using a variety of techniques, such as linear regression, decision trees and random forests. The results showed that random tree-based models outperformed linear models in terms of predicting the market value.
Yaldo and Shamir [9] proposed a methodology to determine the wages of football players depending on their skills. A comprehensive analysis was conducted using 8 different regression methods. Specifically, overpaid and underpaid players were determined using predicted and actual values. Skills that differentiate between overpaid and underpaid players were determined by LDA. Moreover, Relief method was used in order to determine the skills that have the highest impact on salary.

Parath et al [10] introduced a methodology to determine the performance levels of players. In the first stage, an explanatory data analysis was performed using scraping tools. As a result, players were grouped in four fundamental categories in terms of their field positions, attackers, midfielders, defenders and goalkeepers. Each category was individually considered for the determination of player performance level. In the second stage, principal component analysis (PCA) was performed to eliminate adverse impact of noisy features on the learning process. In the last stage, linear regression was individually conducted on the reduced dataset for each specific category. According to the results, the prediction performance of the methodology was promising.

Yigit et al [11] introduced a linear regression model to estimate player values. The regression model was built on the following attributes: ability, age, multiplication of age and concentration, and multiplication of determination and technique. Ridge regression, lasso regression, and tree-based regressors were selected to carry out the regression process. The results showed that the introduced model performed well in terms of predicting a player's value.
Apostolou and Tjortis [12] tried to predict the field position, the number of goals scored for a season and the number of shots during a match. For the performance verification, they selected two top football players and then made comparisons according to their data. However, the generalized performance evaluation scores were not presented in this study; therefore, it is not possible to make analysis concerning the generalization of the proposed method.

There were three objectives Rajesh et al [13] considered: speeding up the selection process of players, facilitating the selection of brand ambassadors, and creating a dream team. It was concluded from statistical analysis and analytical comparisons that a nationality with high potential and higher overall performance was well suited to an international career, and a nationality with moderate potential and performance was more likely to succeed in domestic leagues. Moreover, players between the ages of 21 and 26 were considered to be the most profitable when considering age and performance together. In terms of nationality, England and its neighbors were ranked first, followed by America and Africa.

He et al [14] focused on discovering the relationship between the market value and the player performance. To achieve this, regression models were proposed to predict the market value and the player performance using Lasso regression. While the dataset used in the market value prediction had 100 attributes and 37 forward players who transferred in the same year, the dataset used in the player performance prediction had 40 forward players with 84 attributes. The results showed that each field position brought its own characteristics, resulting significant difference between the players. According to a particular regression model, a good forward player should have the following attributes: few fouls, shots and goals in penalty area, shots on target, goals from out of box, dribble successfully and assists.

Stanojevic and Gyarmati [15] proposed a methodology for data-driven player market value determination. A comparison was made between the Transfermarkt market value estimates (TMVE) and the proposed performance-driven market value estimates (PDMVE). There are three types of features: performance features, player information features and team information features. To select the appropriate hyper-parameters, a small validation set (ten percent of the training data) was used as well as a grid search. The error rate of gradient-boosting trees was slightly lower than that of other methods. Moreover, the top 10 and top 10 undervalued players were listed according to PDMVE. In all metrics, PDMVE achieved a better result than TMVE by 4-7%.

Rao and Shrivastava [16] introduced a methodology to determine the field position of players. In the first stage, data was collected and then a feature selection process was applied. In the second stage, 14 positions were mapped to 3 predefined classes, namely attack, mid and defense. In the last stage, three classifiers, namely multi-layer neural networks, random forest and multinomial logistic regression were performed through GridSearchCV. According to the results, mid and defense positions were generally correctly determined by all three classifiers, while the prediction performance was low for the attack position. The reason is that fewer players play in the attack position. In the case of players with more than one position, the prediction performance was not satisfactory.

Uzochukwu and Enyindah [17] introduced a neural network model to determine characterized attributes of players, involving resistance, speed, technique and physical status. To determine each characterized attribute, some attributes were taken into consideration as inputs for the neural network. After processing neural network models, the average status was calculated by taking the mean of resistance, speed, technique and physical status. The coach will reject the player if the

average status is below 50; if it is 50, the coach can select the player, but he will not be in the starting eleven; if it is above 50, the coach selects the player and he will be in the starting squad. Although neural networks produced promising results in this study, a more comprehensive analysis is required.

Behravan and Razavi [18] introduced a two-stage approach to estimate football players' value. In the first stage, an automatic clustering method based on particle swarm optimization (PSO) was applied to group the data. Accordingly, the data was divided into four groups. Depending on the attributes, the generated groups were goalkeepers with 29 features, strikers with 32 features, midfielders with 28 features and defenders with 30 features. In other words, the proposed method is capable of automatically determine which attributes are important for each cluster. In the second stage, a hybrid version of PSO support vector regression was applied to estimate player values. According to results, PSO performed better than recent evolutionary techniques.

Cotta et al. [19] tried to classify players in order to gain insights into how teams succeed. Firstly, players were divided into three categories: defender, midfielder and attacker. Mean overall attribute for each player was calculated by using 8 years football data from 2007-2014. Since players can change during a season, the top 20 overall players were selected for each category. Using linear regression, a player was predicted to improve or deteriorate for each attribute over the years. And by looking at this data it was seen how attributes change in each team over the years for each category. By following this methodology, they explained one big shocking result in football history, namely Germany's 7-1 win over Brazil in the 2014 FIFA World Cup. In this manner and then by applying PCA and K-means, they were able to explain also one of the biggest success stories in football history, FC Barcelona between 2008 and 2012.

3. CONCLUSION

In this study, we introduce a survey of recent studies on predicting the performance and value of football players. All studies are considered in terms of research purpose, data source, number of features and players, machine learning techniques, evaluation metrics and additional processes. We also summarize the reviewed studies in Table 1 and 2. It can be extracted from Table 1 and 2 that most of the studies are based on regression models, while the remaining studies consider the data as a classification problem. For the subject of estimating player values, these regression studies are understandable. Furthermore, there has been a growing interest in the development of robust sport analysis systems in recent years. To achieve robust football analytic systems, the following issues need to be considered: 1) standard datasets should be provided for researchers to conduct consistent experiments; 2) more comparative studies should be provided to see which machine learning techniques work well in a problem; 3) it is necessary to determine which analyses are crucial for football analytics. Therefore, developing robust sport analysis systems is still an open issue.

In the future, we also plan to develop a methodology using machine learning techniques to estimate player values and categorize players according to their skills.

REFERENCES

[1] Herm S., Callsen-Bracker H.M., Kreis H. (2014). When the crowd evaluates soccer players' market values: Accuracy and evaluation attributes of an online community, Sport Management Review, Vol 17, No:4, p.484-492, DOI: 10.1016/j.smr.2013.12.006

[2] Singh P., Lamba P. 2019). Influence of crowdsourcing, popularity and previous year statistics in market value estimation of football players. Journal of Discrete Mathematical Sciences and Cryptography, Vol 22, No 2, p. 113-126, DOI: 10.1080/09720529.2019.1576333

[3] Kirschstein T., Liebscher S. (2019). Assessing the market values of soccer players – a robust analysis of data from German 1. and 2. Bundesliga, Journal of Applied Statistics, Vol. 46, No 7, p. 1336-1349, DOI: 10.1016/j.ejor.2017.05.005

[4] Müller O., Simons A., Weinmann M.. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. European Journal of Operation Research, Vol. 263, No:2, p.611-624, DOI: 10.1016/j.ejor.2017.05.005

[5] Hanso L., Tama B.A., Cha M. (2022). Prediction of Football Player Value using Bayesian Ensemble Approach. Preprint to be submitted. DOI:10.48550/arXiv.2206.13246

[6] Chavan A., Dondio P. (2019). Recruitment of Suitable Football Player by using Machine Learning Techniques. Msc. Research Project

[7] Al Asadi M.A., Taşdemir Ş. (2021). Empirical Comparisons for Combining Balancing and Feature Selection Strategies for Characterizing Football Players Using FIFA Video Game System. IEEE ACCESS, Vol 9, p. 149266-149286, DOI: 10.1109/ACCESS.2021.3124931

[8] Al Asadi M.A., Taşdemir Ş. (2021). Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques. IEEE ACCESS, Vol 10, p. 22631-22645, DOI: 10.1109/ACCESS.2022.3154767

[9] Yaldo L., Shamir L. (2017). Computational Estimation of Football Player Wages. International Journal of Computer Science in Sport, Vol 16, No 1, p. 18-38, DOI: 10.48550/arXiv.2206.13246
[10] Parath R., Shah S., Surve A., Mittal J. (2018). Player Performance Prediction in Football Game, 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA2018), p. 1148-1153, DOI: 10.1109/ICECA.2018.8474750
[11] Yiğit A.T., Samak B., Kaya T. (2020). Football Player Value Assessment Using Machine Learning Techniques, Springer Nature Switzerland, p. 289-297, DOI: 10.1007/978-3-030-23756-1_36
[12] Apostolou K., Tjortjis C. (2019). Sports Analytics algorithms for performance prediction 10th International Conference on Information, Intelligence, Systems and Applications (IISA), DOI:10.1109/IISA.2019.8900754
[13] Rajesh P., Alam M., Tahernezhadi M., (2020). A Data Science Approach to Football Team Player Selection. IEEE International Conference on Electro Information Technology (EIT), p. 175-183, DOI:10.1109/EIT48999.2020.9208331
[14] He M., Cachucho R., Knobbe A. (2015). Football Player's Performance and Market Value. Proceedings of the 2nd workshop of sports analytics, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)
[15] Stanojevic R., Gyarmati L. (2016). Towards data-driven football player assessment. IEEE 16th International Conference on Data Mining Workshops, p. 167-172, DOI: 10.1109/ICDMW.2016.0031
[16] Rao V., Shrivastava A. (2017). Team Strategizing using a Machine Learning Approach. International Conference on Inventive Computing and Informatics (ICICI 2017), p. 1032-1035, DOI: 10.1109/ICICI.2017.8365296
[17] Uzochukwu O. C., Enyindah P. (2015). A Machine Learning Application for Football Players' Selection. International Journal of Engineering Research & Technology (IJERT). Vol. 4, No 10, p.459-465, DOI : 10.17577/IJERTV4IS100323
[18] Behravan I., Razavi S.M. (2020). A novel machine learning method for estimating football players' value in the transfer market. Soft Computing, Vol. 25, p 2499-2511,DOI:10.1007/s00500-020-05319-3
[19] Cotta L., Benevenuto F., Vaz de Melo P., Loureiro A. (2016). Using FIFA Soccer video game data for soccer analytics, Workshop on large scale sports analytics, DOI: 10.1145/1235

Techno-Science *Paper ID: 1215552*