4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15067

Influence of popularity on the transfer fees of football players

Pilar Malagón-Selma[1], Ana Debón[1], Josep Domenech[2]
[1]Centro de Gestión de la Calidad y del Cambio, Universitat Politècnica de València, Spain,
[2]Departamento de Economía y Ciencias Sociales, Universitat Politècnica de València, Spain.

Abstract

Search popularity, as reported by Google Trends, has previously been demonstrated to be useful when studying many time series. However, its use in cross-section studies is not straightforward because search popularity is not provided in absolute terms but as a normalized index that impedes comparisons. This paper proposes a novel methodology for calculating popularity indicators obtained from Google Trends to improve the prediction of football players' transfer fees. The database is formed by 1428 players who competed in LaLiga, Premier League, Bundesliga, Serie A, and Ligue 1 on the 2018-2019 season. Random forest algorithm and multiple linear regression are used to study the popularity indicators' importance and significativity, respectively. Results showed that the proposed popularity indicators provide significant information to predict players' transfer fees, as models including such popularity indicators had lower prediction error than those without them. This study's developed method could be used not only for analysts specialized in sports data analysis but for researchers of other fields.

Keywords: *Popularity Indicators; Google Trends; Transfer fees*

1. Introduction

With 158 years of history, football is not only the King sport of today's society but one of the most profitable businesses in the world. According to Ajadi et al. (2021), the combined turnover of the top 20 clubs was €8.2 billion in 2019/20. However, such amounts of income are accompanied by significant expenses. In 2017, Paris Saint-Germain F.C. carried out the most expensive transfer in history, paying €222 million to F.C. Barcelona for Neymar Jr. A year later, this same team bought Kylian Mbappé for €180 million, becoming the second most expensive transfer fee[1] in the history of this sport (Trujillo, 2021). These expenses can only be understood by considering that the main assets of football teams are the players. Thus, given the impact of transfer fees on the economy of football clubs, academics, managers, and other experts have tried to find their main determinants. Factors affecting the transfer fees include the players' performance, position (forward, midfielder, defender, or goalkeeper), the club they play for and, physical characteristics (height, age, etc.) (Garcia-del-Barrio & Pujol, 2007; Herm, Callsen-Bracker, & Kreis, 2014; Müller, Simons, & Weinmann, 2017).

Furthermore, football players are brands themselves, and they have been benefited from the emergence of social networks such as Instagram or Twitter. So, it seems reasonable to study their online popularity and how it impacts the transfer fees, especially if this information is open and easy to access. Previous research has already used popularity measures, such as the followers on social media (Müller et al., 2017; Hofmann, Schnittka, Johnen, & Kottemann, 2019) and their exposition in Google, measured as the number of hits (Garcia-del-Barrio & Pujol, 2007; Herm et al., 2014; Hofmann et al., 2019) to predict the football player transfer fees. Müller et al. (2017) also incorporate Reddit posts, Wikipedia views, YouTube videos, and a Google Trends search index[2]. In this regard, Garcia-del-Barrio and Pujol (2007) and Herm et al. (2014) have found the number of hits in Google results is statistically significant in predicting players' transfer fees, while Hofmann et al. (2019) did the same for the number of followers on social media. Similarly, Müller et al. (2017) found all popularity variables to be statistically significant except the Google Trends search index. This could be because GT does not provide time series of absolute searches but term-dependent normalized indexes from 0 to 100, so they cannot be directly used to compare different players.

This article proposes novel ways to use GT to measure player popularity by requesting several terms (i.e., player names) simultaneously. To demonstrate its usefulness, this

[1] Actual prices paid on the market (Müller et al., 2017).

[2] Google Trends is a tool that allows users to measure the interest that a topic or a person arouses in the world over time according to the number of searches in Google Search Engine (Rogers, 2016).

methodology has been applied to help predict the transfer fees of the players sold during the summer market of the 2018-2019 season.

The rest of the paper is organized as follows. The second section is devoted to explaining how the proposed popularity indicators have been calculated. The third section describes the database and the statistical methods used for carrying out the analysis. The fourth and fifth sections introduce the results obtained and the conclusions achieved, respectively.

2. Popularity indicators with Google Trends

The time series provided by Google Trends (Rogers, 2016) contain a relative index of term popularity, normalized from 0 to 100, which takes value 100 in the period with the highest number of searches. This normalization makes it difficult to compare player popularities since the corresponding series are individually normalized. That is, all series are rescaled considering their maximum. According to Rogers (2016), one way to put the search interest into perspective is to add additional terms. Thus, using two terms (each one representing a player) simultaneously, the results of both series are jointly normalized, i.e., with respect to the highest popularity of any of the terms. Therefore, both series are on the same scale, and it is possible to compare them.

Unfortunately, GT series are reported as whole numbers instead of real numbers. Thus, if a famous player is compared to an unpopular one, GT reports a search index of 0 for the latter, making it difficult to compare the popularity of less searched players. To deal with this issue, we propose to use different reference players according to the relative popularity and position, since the notoriety of a player depends on his position.

In this study, three popularity layers were defined (“High” for the most popular players, “Middle” for the relatively popular, and “Low” for the less popular), each one with a specific reference player. The reference player of the first layer was the one who, compared to the rest, had the highest average search index (in the case of the forwards, Cristiano Ronaldo). The reference player for the second layer was a player whose average popularity index was 1 when put together with the reference player in the first layer. Among all those satisfying this criterium, the least popular one was selected as the reference for the second layer. Less popular players (that is, receiving an average search index of 0 when compared to the layer-1 reference player) were then compared to the reference player of the second layer. This process was repeated in the three levels of the three considered player positions (defender, midfielder and forward). After that, all series were rescaled to account for the different reference players used.

Once the time series of the weekly popularity of the players are on the same scale, their information is summarized in six popularity indicators that can be used in cross-sectional studies: First Principal Component (CP1), mean, median, maximum, minimum and variance.

3. Methodology

The following section presents the database used to carry out the study and the statistical methods used in the predictive analysis. Free R software was used for the analysis (R Core Team, 2019).

3.1. Models

In order to know if the proposed indicators have a significant impact on the transfer fees prediction, two different models were considered. In Model 1, considered as the baseline, the transfer fee for each player *i* is explained by his characteristics[3] and his performance[4].

$$\text{Transfer fee}_i = f(characteristics_i, performance_i) \tag{1}$$

Model 2 extends Model 1 by including the popularity indicators described in Section 2.

$$Transfer\ fee_i = f(characteristics_i, performance_i, popularity_i) \tag{2}$$

3.2. Data

The database used to carry out the analysis was formed by 1428 players who competed in LaLiga, Premier League, Bundesliga, Serie A and Ligue 1 on the 2018-2019 season with 36 explanatory variables related to player characteristics, performance, and six popularity indicators[5]. To train the models, the estimated market value[6] of 1235 players not sold where used. The model error was assessed using the transfer fees of the 193 players sold during the summer market after that season.

3.3. Methods

Random Forest algorithm (RF) (Breiman, 2001) and Multiple Linear Regression (MLR) (Berry, Feldman, & Stanley Feldman, 1985) where used to fit the models. Before carrying

[3] Player characteristics: Position, age, height, and contract.

[4] Player performance: Playing time, aerial duels accuracy, tackles accuracy, interceptions, shots intercepted, fouls, yellow cards, red cards, goals, shots, shots accuracy, assists, dribbles, crosses, corners, passing accuracy, short passes accuracy, long passes accuracy, key passes, progressive passes, deep passes, penalty area, last half quarter, and free kicks.

[5] Popularity indicators were calculated using values for the time period from 17 May 2018 to 26 May 2019 (popularity per week).

[6] Amount of money that a club would be willing to pay for an athlete to sign a contract, regardless of an actual transaction (Herm et al., 2014). Source: www.transfermarkt.com

out the MLR and for alleviating the multicollinearity, variance inflation factors (VIF) were obtained using the `vif_function` (Thompson, 2013), removing those variables with VIF >5 only for MLR. Later, the most relevant variables were selected in the fitted linear model according to the Akaike information criterion (AIC) (Akaike, 1974) using the `MASS` R-package (Venables & Ripley, 2002).

In addition, the repeated k-fold cross-validation technique (in this case, k=5 and repetitions=5) was used to optimise the hyperparameters of the training set for both methods, RF and MLR, using the `caret` R-package (Kuhn, 2020). Finally, the model's performance was obtained on transfer fees of 193 players, who had not been used to build the model.

4. Results

After applying the methodology Table 1 shows the performance for each method and model measured through the root mean square error (RMSE).

Table 1. Summary of model performance measured by means of the RMSE (EUR).

	RF	**MLR**
Model 1 (baseline)	16,583,803	17,045,285
Model 2 (popularity)	12,083,185	15,338,117

Source: Own calculations

According to Table 1, using the popularity indicators, the RMSE decreased by €1,707,168 and €4,500,618 for the MLR and RF methods, respectively.

Table 2. Variables selected by the Multiple Linear Regression after applying the AIC in model 2

Type of variables	**Variables**
Player characteristics	Position, age, and contract
Player performance	Playing time, aerial duel accuracy, fouls, goals, shots, assists, dribbles, short passes accuracy, passes in the last quarter of the opponent half, deep passes, free kicks, and corners
Popularity indicators	Variance, minimum, median

Source: Own calculations

Table 2 shows variables selected by the MLR after applying the AIC in model 2. Note that, after applying the *vif_function* the variance, minimum, and median were the only popularity indicators that remained in the model. Thus, the MLR selected these three popularity indicators included in the model.

RF algorithm allows knowing the importance of the variables in the regression model. Liaw and Wiener (2002) incorporated, in the randomForest R package, the calculation of the average increase of the mean squared error (IncMSE%) in the out-of-bag when one variable's values are permuted in the training dataset while the others remain unchanged (the greater the prediction error, the greater the importance of the variable). Figure 1 shows the importance of the variables according to the IncMSE% in the model 2.

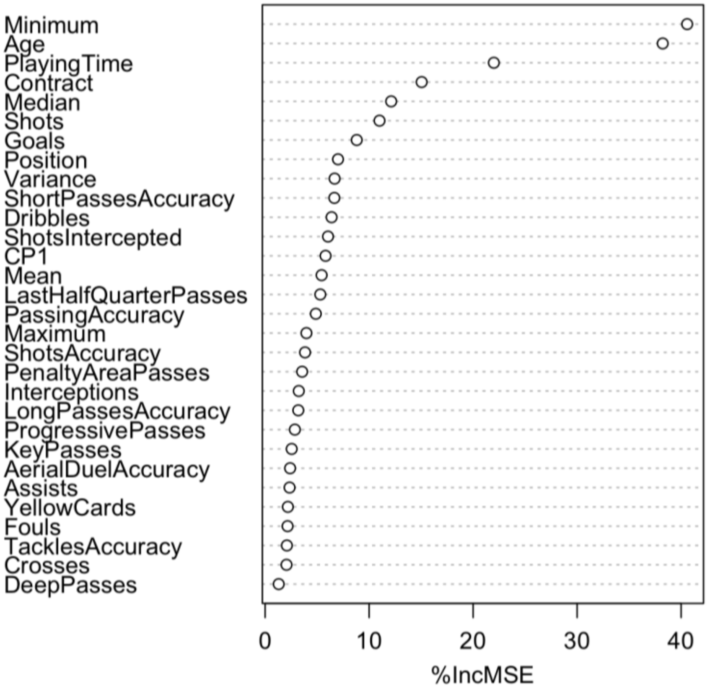

Figure 1. Importance of variables of Random Forest algorithm for the model 2. Source:Own calculations.

Figure 1 shows that in the case of RF algorithm, the most important variable is the "Minimum" popularity indicator, which stores information about the week in which players were least searched. Additionally, in the same way as MLR, the variables "Median" and "Variance" take a relevant position.

4. Conclusion

This work proposed new ways to use GT data to measure player popularity for predicting his corresponding transfer fee. First, because the time series given by GT is individually normalized, it should not be used directly to measure player popularity. Thus, this document recommends using reference players classified by popularity levels as a possible solution. Second, the results (Table 2 and Figure 1) show that the popularity indicators calculated through the proposed methodology (see section 2) improve the prediction of transfer fees. This information may be helpful to analysts who might add these indicators to their models to improve transfer fees prediction.

Acknowledgments

This work was partially supported by grants PID2019-107765RB-I00 and funded by MCIN/AEI/10.13039/501100011033.

References

Ajadi, T.; Bridge, T.; Hanson, C.; Hammond, T.; Udwadia, Z. (2021). Deloitte Football Money League 202. *Deloitte Sports Business Group,* 2-58. https://www2.deloitte.com/uk/en/pages/sports-business-group/articles/deloitte-football-money-league.html.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19(6), 716-723.*

Berry, W., Feldman, S., & Stanley Feldman, D. (1985). *Multiple regression in practice.* Thousand Oaks, CA: Sage.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Garcia-del-Barrio, P., & Pujol, F. (2007). Hidden monopsony rents in winner-take-all markets?sport and economic contribution of Spanish soccer players. *Managerial and Decision Economics, 28*(1), 57-70.

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, 17*(4), 484-492.

Hofmann, J., Schnittka, O., Johnen, M., & Kottemann, P. (2019). Talent or popularity: What drives market value and brand image for human brands? *Journal of Business Research, 124,* 748-758.

Kuhn, M. (2020). *Caret: Classification and regression training.R package version 6.0-86.*

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2*(3), 18-22. http://CRAN.R-project.org/doc/Rnews/.

Müller, O., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. *European Journal of Operational Research, 263*(2), 611-624.

R Core Team. (2019). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Rogers, S. (2016, September 1). *What is Google Trends data — and what does it mean?* (Google News Lab) Retrieved December 12, 2021, from GoogleNews2016: https://medium.com/google-news-lab/what-is-google-trends-data-and-what-does-it-mean-b48f07342ee8.

Thompson, S. (2013, February 5). *Collinearity and stepwise VIF selection*. Retrieved October 15, 2020, from http://beckmw.wordpress.com/2013/02/05/collinearity-and-stepwise-vif-selection/.

Trujillo, I. (2021, August 27). *¿Qué lugar ocupará Mbappé entre los fichajes más caros de la historia del fútbol?*. Retreived March 9, 2021, from LaRazón: https://www.larazon.es/deportes/futbol/realmadrid/20210827/fz345lazmreqjfsn5wrq4u5ogy.html.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S.* New York: Springer.