Received: 2 January 2019 | Revised: 29 April 2019 | Accepted: 29 April 2019
DOI: 10.1002/cpe.5353

WILEY

SPECIAL ISSUE PAPER

Data-driven exploratory approach on player valuation in football transfer market

Yunhu Kim | **Khac-Hoai Nam Bui** | **Jason J. Jung**

Department of Computer Engineering, Chung-Ang University, Seoul, Korea

Correspondence
Jason J. Jung, Department of Computer Engineering, Chung-Ang University, Seoul 156-756, Korea.
Email: j2jung@gmail.com

Present Address
Jason J. Jung, Chung-Ang University, 84 Heuk-seok Ro, Dong-jak Gu, Seoul 156-756, Korea.

Funding information
Chung-Ang University Research Scholarship Grants; National Research Foundation of Korea, Grant/Award Number: NRF-2017R1A2B4010774

Summary

Transfer markets in football have attracted the interest of researchers in economy and management. In this paper, we propose a high level analysis approach for classifying player valuation based on their performance during recent seasons. In particular, several data analysis techniques such as regression analysis, feature selection, and cluster analysis are presented for classifying players in term of performances and transfer fee. Specifically, by collecting and analyzing data from *Wholescored*, the largest detailed football statistics website, we have defined players into four groups, which include (1) Low performance and low transfer fee (LPLF), (2) Low performance and high transfer fee (LPHF), (3) high performance and high transfer fee (HPHF), and (4) high performance and low transfer fee (HPLF). The results in the implementation section show that, with the differences positions, there are different required skills that affect to the performance of players. We expect that this study can contribute to the management of Football Teams in terms of integrating these analyses into their management strategy.

KEYWORDS

data analytics, feature selection, football player transfer market, regression analysis, sport big data

1 | INTRODUCTION

As Big Data Analysis becomes an issue, it is closely related to sports field.[1] In the field of baseball, paradigm has changed with the emergence of Sabermetrics, which already use statistical methods.[2] In the football field, data analysis has applied later than in the baseball field because the football data is difficult to explain statistically.[3] The research in existing football data focuses on Social Network Service (SNS) analysis,[4-6] winner prediction,[7] and so on. However, there is needed for taking into account on the transfer market of football players since the scale of the football transfer market is growing steeply.[8] Table 1 depicts the most expensive players who have been transferred by previous seasons. As a result, the highest transfer fee of each player has been renewed every year, and the result shows that the size of the football transfer market is increasing.

The transfer is closely related to the ranking of the football club. If a football club bought a good player, it would rank high and the high ranking would have made a lot of money,[9] for instance, the prize money, broadcasting rights, player transfer fee, and so on.[10] However, the problem of the current transfer system in the transfer market is determined by the intuition of the coach and the board members, not the objective basis.[11] Specifically, the current transfer system is processed as follows.

- The first step is dispatching a scouts to observe the play of rookies.
- After reporting to the coach and board members about the observed player, they approve the proceeding of negotiation.
- The club offers a transfer to the player, officially.
- Down payment and salary are taken into account.
- Transfered player's physical health will be checked through a medical test.
- Finally, the players complete the transfer.

For instance, in 2006, Chelsea Football Club (FC) paid AC Millan striker Andriy Shevchenko for astronomical money. Despite the fact that Andriy Shevchenko's scoring ability has been reduced a little bit before the transfer, the transfer was made since the owner Roman Abramovich

TABLE 1 The most expensive player in each year from 2012 to 2018

Season	Name	Transfer Fee	Salary	Age	Previous Team	Next Team
2012/2013	Thiago Silva	50M	9.8M	28	AC Milan	Paris Saint-Germain
2013/2014	Gareth Bale	91M	7M	24	Tottenham	Real Madrid
2014/2015	Luis Suarez	94M	12.5M	27	Liverpool	FC Barcelona
2015/2016	Kevin De Bruyne	75M	14.6M	24	Wolfsburg	Manchester City
2016/2017	Paul Pogba	105M	19.1M	23	Juventus	Manchester United
2017/2018	Neymar	222M	36.4M	25	Barcelona	Paris Saint-Germain
2018/2019	Philippe Coutinho	120M	12M	26	Liverpool	FC Barcelona

of Chelsea FC want to transfer the player.[12] As a result, after the transfer to Chelsea FC, Andriy Shevchenko has been listed as one of the worst signings in Chelsea FC's history, with only nine goals in 58 games.[13]

In this regard, we propose a novel approach based on data analysis for football transfer market. To the best of our knowledge, this is the first study in this research area. In particular, we have collected actual observed data from *Whoscored*,* which is regarded as a dataset for analyzing. Based on the collected football player data, we take into account predicting the performance after transferring and clustering by comparing to the transfer fee. In addition, feature selection techniques for each position have been used to extract important player skills and then clustering the following the performance and the transfer fee of players. Specifically, the main contributions of this study are described as follows.

- We propose a novel approach of football data analysis focusing on the performances and transfer market of football players. Moreover, it is objective and highly reliable because it has introduced to the research using the real observed data of the players.
- Second, this paper can contribute to maximizing performance with minimal investment in each club. Although clubs have different policies, the ultimate goal of all clubs is top performance.
- Based on the results of this study, more active research studies can be expected in terms of applying data analysis of football field.

The rest of this paper is organized as follows. In Section 2, the background of the sport industry is introduced. Moreover, in this section, we present some techniques of data analysis that have been used in this study. In Section 3, we propose an attribute-based Feature Selection algorithm for Football Player Transfer Market which is the main contribution of the paper. In Section 4, we report the experimental results based on the dataset of players. In this regard, we take English Premier League (EPL) into account as the case study. Some discussions and future works are concluded in Section 5.

2 | BACKGROUND

2.1 | Data analysis for sport industries

Nowadays, we are living in the era of big data that extracting values through a huge amount of data.[14,15] Big data analysis is used in various fields such as politics, society, economy, culture, and so on.[16,17] For instance, in the field of baseball, they used Sabermetrics. Sabermetrics is a way to make an objective assessment of the players based on their records. A study of multiple regression analysis using Conventional and Sebermetrics Measures as the variables was also conducted by extending the concept of Sabermetrics. Beneventano et al[18] proposed a Runs-Scored model to calculate the weights of on-run on offense and run prevention on pitching and defense. This model provided an opportunity to discover undervalued players by calculating the performance of players in the team.

In field of football, Germany National Football Team (GNFT) have used *SAP Match insight*; the solution analyzes the team's performance and the performance of its opponents to help GNFT win the 2014 Brazil World Cup.[19] Particularly, SAP Match Insight attaches a wearable Sensor and collects respiratory, pulse, and activity data of the players through wearable Sensors. Specifically, Data has collected at one training per players which is able to collected more than 12 000 per minute. The collected data then has sent to the wireless receiver installed in the training area and the large amount of data transmitted is assisted by the analysis platform "SAP HANA."†

In particular, methods for evaluating players and teams based on data have been studied in the Data analysis for football. The national football team's ranking was measured by FIFA Ranking. However, the problem of the weight between the low-ranking team and the high-ranking team and the problem of home advantage were not considered. To complement the problem, World Football Elo Ratings (WFER), which uses the Elo System, a chess player evaluation system, was studied.[20] Specifically, WFER is one of the ways to rank the players, according to the national football team's records.[21-23] Particularly, the score is measured by the points which determined by both teams and the weight is determined by winning or losing.

* https://www.whoscored.com
† https://www.sap.com/

There are many studies of the team's evaluation, but not many of the individual players' performance evaluations. To complement this, Sæbø and Hvattum have introduced the Plus-Minus player statistics measure.[24] Particularly, Plus-Minus player statistics is a concept to consider the level of teammates and the level of opponents. In order to apply this idea to the football field, each game was divided into a set of segments, considering a collection of past games rather than a simple rating. They used regression analysis of the calculated rating and transfer fee values to give objective player ratings. As a result, it has been proven that players with some characteristics require too high a transfer fee for their ability. However, this study has a limitation that only rating value is used instead of detailed records.

In this paper, instead of rating values of players, we collected detailed records of the last three seasons before moving the team, for instance, Kevin De Bruyne's Goal, Assist, Shoot, Dribble, Pass Accuracy, Key pass, and so on. Regression analysis of the collected data predicted the performance of the player after the transfer. Then, clustering techniques are applied for the players who have completed the transfer to study the range of ransom value of the current players. Finally, we conduct clustering of the players who completed the transfer to predict the range of current transfer fee of the players.

2.2 | Cluster analysis

Machine learning, which is a method of learning computers using data, is divided into supervised learning and unsupervised learning.[25] Specifically, Supervised learning is a way of learning computers with labels given to data. On the other hand, Unsupervised learning is a way of learning computers without labeling data.[26] Since the football data used in this paper uses record real data of players, there is no label which is given for each data. Therefore, we used cluster analysis, one of the unsupervised learning, to characterize the players. The methods of cluster analysis are partitioning, $\mathcal{K}$-means, and so on.[27]

Specifically, we used $\mathcal{K}$-means, which has been used as an exploratory approach because it is an easy structure and applicable to various data. Technically, the $\mathcal{K}$-means algorithm defines the number of clusters $\mathcal{K}$ and chooses Centroid of the initial $\mathcal{K}$ clusters. Each data is then assigned to the closest centroid cluster. The next step is to calculate the centroid of the new cluster and re-classify the distance-based clusters back to the calculated centroid until cluster is unchanged.[28]

2.3 | Feature selection

Feature Selection is a well-known technique that using a large number of variables in clustering than optimal variables, which reduces prediction accuracy.[29] For this reason, extracting useful information from high-dimensional data requires the use of statistical techniques to reduce noise and redundant data. By using feature selection method, we can improve the computation speed and reduce the complexity of the model comparing the case of using high dimensional original football player records. Technically, Feature selection is simply divided into three methods. (i) The first method is a Filter method that filters out variables before applying them to the model, ranks them using statistical indicators such as correlation coefficients, and excludes those with low influences; (2) the second is the Wrapper method, which selects the combination with the highest score among the completed models by learning which different combinations of variables exist; (3) the final method is the Embedded method, which combines the Filter method and the Wrapper method. This method is a method for selecting a Wrapper method by selecting optimized variables in the learning process.

In this paper, we used Boruta algorithm, which belongs to the Wrapper methods.[30] Particularly, the Boruta algorithm is based on Z score with random forest which maximizes diversity and has a considerably superior predictive power. Since it makes decisions based on the prediction results of many trees, it has high predictive power and ensures stability. The basic idea of this algorithm is to shuffle the predicted variable values, combine them with the original predicted variables, and then create a random forest on the merged data set. Then, by comparing the original and random variables to measure the importance of the variables, we only selected variables that are more important than the random variables.

3 | EXPLORATORY APPROACH FOR FOOTBALL PLAYERS TRANSFER MARKET

Figure 1 depicts the system architecture of the proposed approach. Specifically, our approach is divided into three main steps as follows.

- The first step is to collect data by position from Whoscored, a football professional website. In football, positions are divided into four categories. Supporting Forward (FW) is a position that plays near the opponent goalpost and scores a goal when the opportunity comes. Defender (DF) is a position that plays a role in obstructing the opposing team from scoring. The Midfielder (MF) is the role of playing and position between the FW and DF, with detailed roles divided according to the location and tactics. Goalkeeper (GK) is the only position in the penalty box where they can handle the ball with their hands, and plays the role of defending the goalpost. In detail, the FW is divided into Wing Forward (WF) and Central Forward (CF). MF is divided into Attacking Midfielder (AMF), Wing Midfielder (WM), Central Midfielder (CM) and Defensive Midfielder (DMF). DF is split into Full Back (FB) and Center Back (CB).
- The second step is to perform a regression analysis based on the player's before the transfer data and calculate the predicted value of the performance after the transfer.

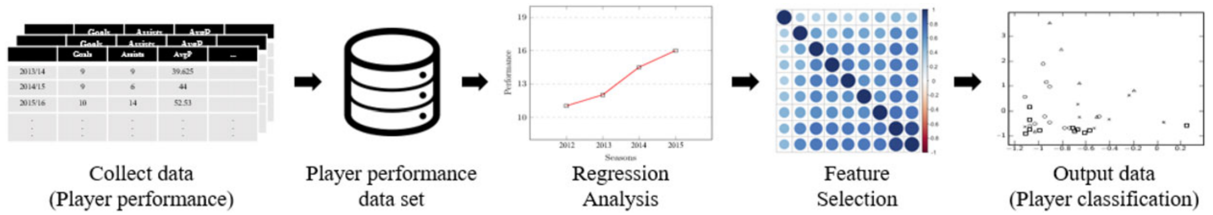

FIGURE 1 The process for extracting classification features of players

TABLE 2 Angel Di Maria's 3 years records before transfer

Season / records	Assist	Key pass	Dribble
2012	10	1.95	1.3
2013	22	2.13	2.7
2014	12	2.05	1.95

- The final step is feature selection for each position using a Boruta algorithm. Then, k-means clustering is performed on the transfer fee using only selected variables. It is possible to explain the value of the athlete through the characteristics of the player in each cluster.

3.1 | Regression analysis

In this paper, we apply regression analysis based on the data of the transferred players to predict the performance of the player after the transfer. We denote a transferred player, the year before transfer, and the transferred position as $p_m \in \mathcal{P}, m \in [1, M]$, $\tau_l \in \mathcal{T}, l \in [1, L]$, and $\rho_c \in \mathcal{R}, c \in [1, C]$, respectively.

When a player's records is $s_n \in \mathcal{S}, n \in [1, N]$, the predicted performance of each player can be calculated as follows:

$$p_m(\tau_l, \rho_c) = (s_1, \dots, s_n). \tag{1}$$

Supporting τ_l is the value of the predicted performance of player in each season l by using the regression analysis, which is able to be estimated as follows:

$$t\hat{a}u_l = \beta_0 + \beta_{s_i} p_m, \tag{2}$$

where β_0 and β_{s_i} denote for the regression interception and the weight of each skill, respectively.

By calculating the regression line through each record, the records of the players after the transfer can be predicted, eg, the three-year record of Angel Di Maria, a player in AMF position, who moved from 2015/16 Season Manchester United to Paris Saint-Germain. The regression analysis is performed by applying the AMF's representative record of assist and key passes per competition and the average number of dribbles per competition at s_n. The data used in the examples is shown in Table 2. By estimating the regression line based on the data, Angel Di Maria's predicted record for the 2015 season is 16.67 assist, 2.14 key pass, and 2.63 dribble, respectively.

3.2 | Boruta algorithm-based extracting important features of football skills

In this paper, we used the Boruta algorithm Which is based on Random Forest algorithm. The importance of a feature is calculated as the loss of classification accuracy resulting from any sequence of features. Specifically, the calculation for each tree in the random forest has been used for the classification. Technically, random forest algorithm does not utilize Z scores with mean and standard deviations since the distribution does not follow the normal distribution. Therefore, it is not directly related to the statistical significance of the importance of the features. On the other hand, the Boruta algorithm uses Z score to measure the importance of considering the fluctuation of average accuracy loss in a tree in a random forest. Specifically, the process of the algorithm is described as follows.

- For the first step, we insert the predicted features (eg, Dribble, Goals, Assists, and so on) for each player made by regression in to original data.
- Second step is to duplicate and permute the data, then run a Random Forest on the Shadow Feature to find Z score. The maximum value of Z Score is defined as MSZ. We also run a random forest on each of the existing data features to get Predicted performance's Z Score. Regarding the values of Predicted Performance's Z Score and MSZ, in case of the Predicted performance's Z Score is bigger than MSZ, the number of hit will be increased.
- On the other hand, in case of the Predicted Performance's Z score is smaller than MSZ, performing a two-side equality test on the MSZ, and then inserting it into the Importance variable. Then, if the Importance variable is greater than the MSZ at a statistically significant level, the feature value will be inserted into the Output value which includes the set of important values.

3.3 | Attribute-based feature selection for football player transfer market

Algorithm 1 depicts the process of our proposed approach which includes five main steps as follows. (1) In the first step, we collected data from Whoscored to pre-processing; (2) regression analysis for prediction player performance is proposed in second step for predicting the performance of players in the next year; (3) after classifying players in each positions in step 3, we apply Attribute-based Feature Selection for each positions to extracting important values (skills) of each positions in step 4; and (4) finally, in step 5, we use K-means for Player Classifications in terms of performance and transfer fee.

Algorithm 1 Attribute-based feature selection for football player transfer market

```
Input : Football player dataset
Set of Player 𝒫 ;   Set of Skills 𝒮 ;   Set of Positions ℛ;
  Set of Weighting β;   Set of Seasons tâu; ;   Set of Clusters 𝒞ℒ
Output : Player Classification for Next Season
Initialization;
// Regression Analysis for Prediction the next Season //
while each tâu do
    for each s ∈ 𝒮 do
        for each ρ ∈ 𝒫 do    tâu_l = β_0 + β_{s_l} ρ;
        end for
    end for
end while
// Attribute-based Feature Selection //
for each r ∈ ℛ do
    for each ρ ∈ 𝒫 do
        if ρ_c == r then    r ← ρ_c ;
        end if
    end for
end for
// Player Classification //
for each r ∈ ℛ  do
𝒫𝒱_r := Return Important − Feature ;
end for
while ρ ∈ 𝒫 do
    for each i ∈ 𝒫𝒱 do    //
        Initialvalue = ∅;    Cluster = ∅ ;
        for Centroid c do Dist = Distance(c)
            if Dist < Initialvalue then    Initialvalue = Dist ;    Cluster := c
            end if
        end for
    end for
    RecalculateCentroids(c) ;
until convergence
end while
```

4 | EXPERIMENTATION

In this section, we collected the data to implement our approach, which is described in Section 3. Although the transfer fee is enclosed by the competition between the clubs, the dataset of the players who completed the transfer were collected in the EPL using the transfer fee list released by the British public service broadcaster BBC (British Broadcasting Corporation) Sports.

In this regard, by using the history data of *Whoscored*, which is an open website of football experts to measure ratings based on the records of each match, we collected the data of the three years before transferring the players who completed the transfer. In this experimentation, *GK*, a defensive position was excluded since *whoscored* does not provide serviced skills of players in the position. Specifically, the detailed description of variables, which are provided in *Whoscored*, is presented in Table 3.

TABLE 3 Description of variables

Skills	Description	Skills	Description
Apps	Number of games	Crosses	Crosses per game
Mins	Minutes played	LongB	Long balls per game
Goals	Total goals	ThrB	Through balls per game
Assists	Total assists	Tackles	Tackles per game
SpG	Shots per game	Inter	Interceptions per game
Drb	Dribbles per game	Fouls	Fouls per game
Fouled	Fouled per game	Offsides	Offside won per game
Off	Offsides per game	Clear	Clearances per game
Disp	Dispossessed per game	DrbP	Dribbled past per game
UnsTch	Bad control per game	Blocks	Outfielder Blocks per game
KeyP	Key passes per game	OwnG	Own goals
AvgP	Passes per game	Rating	Evaluated football player by expert
PS	Pass success percentage	Fee	Transfer fee

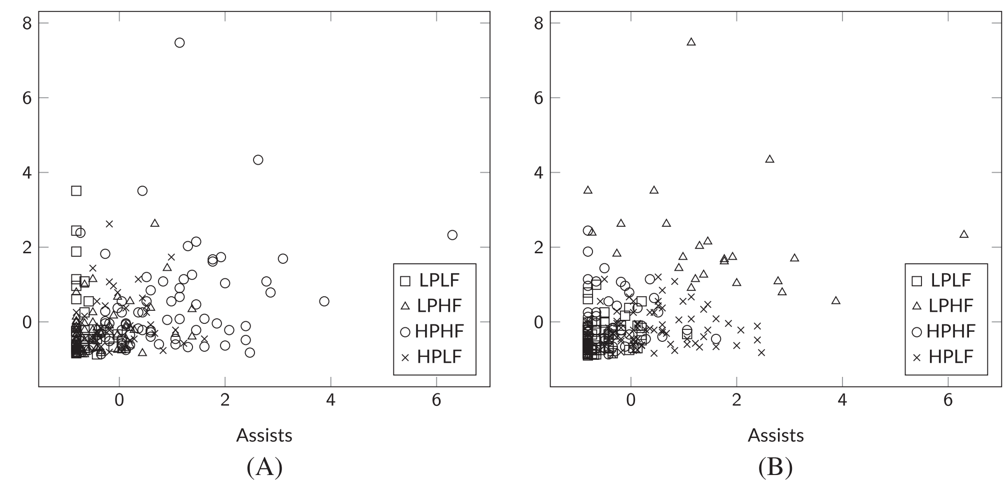

FIGURE 2 All players Clustering (ylabel = *Fee*). A, All players; B, All players with feature selection

Although each club has different policies, they expect to be ranked higher due to the transfer of players. In other words, they want players to record the best performance in proportion to the cost spent in the club. For example, Chelsea FC, who has signed Alvaro Morata for 60 million euros, would have transferred him with his hopes that he would score at least 15 goals in a season. However, as a result, he did not satisfy Chelsea FC's expectations. For the club's ranking, it would have been better if the coach had transferred different types of players through clustering based on data. In this paper, when using *NbClust* clustering, the number of cluster*K* in the cluster used an *K-means* package that provides the best number of clustering results out of all combinations of *R*.[31] Since each variable has difference sizes, scaling was performed before implementing *K-means* clustering. Moreover, since we clustered by using all 26 variables, the values for Assists, AvgP, Goals, and variables are visualized as representative of 26-dimensional data.

Figure 2A shows the result of clustering the transfer fee of all collected data. We could see that clustering was not performed clearly for each cluster. Figure 2B is obtained by performing feature selection for all players and clustering the same method using selected skills.

By using the Boruta Algorithm for extracting Important Features of Football Skills, Table 4 depicts results of extracting the important import skills that affect the transfer fee for each position. By observing, we are able to confirm that the selected skills for the forward, midfielder, and defender categories were similar.

Thereby, Figure 3 shows the result of performing the Attribute-based Feature Selection for the Football Player Transfer Market for one position in the center by forward, midfield and defender categories. Since the selected skills for each position are different, we showed clustering results for the three selected skills. Figures 3A to 3C are the result values for *CF* position. On the other hand, Figures 3D to 3F are the result values for *CM* position.

Moreover, *LPLF* label interpreted it as a cluster of rookie who showed low transfer fee and low performance. The *LPHF* label represents a cluster of players that show lower performance than high transfer fee. The *HPHF* label means a group of players at a reasonable transfer fee that shows high performance at a high price, and the *HPLF* label is a group of efficient players that show low transfer fee and high performance.

TABLE 4 Selected skills with feature selection

Position	Seleted Skills
CF	Assists, AvgP, Clear, Drb, Disp, Fouled, Fouls, Goals, KeyP, Mins
WF	Assists, AvgP, Crosses, Drb, Disp, Goals, Tackles, KeyP
AMF	Assists, AvgP, Blocks, DrbP, Goals, Tackles, KeyP, Mins
CM	Assists, AvgP, Drb, Fouled, Fouls, Inter, KeyP, Mins
DMF	AvgP, Drb, DrbP, Tackles, Inter, Mins
WM	AvgP, Crosses, Drb, Tackles, KeyP, PS
FB	Apps, Assists, Crosses, DrbP, Tackles, KeyP, Mins
CB	Apps, Block, Clear, Crosses, Drb, Goals, Inter, KeyP, Mins

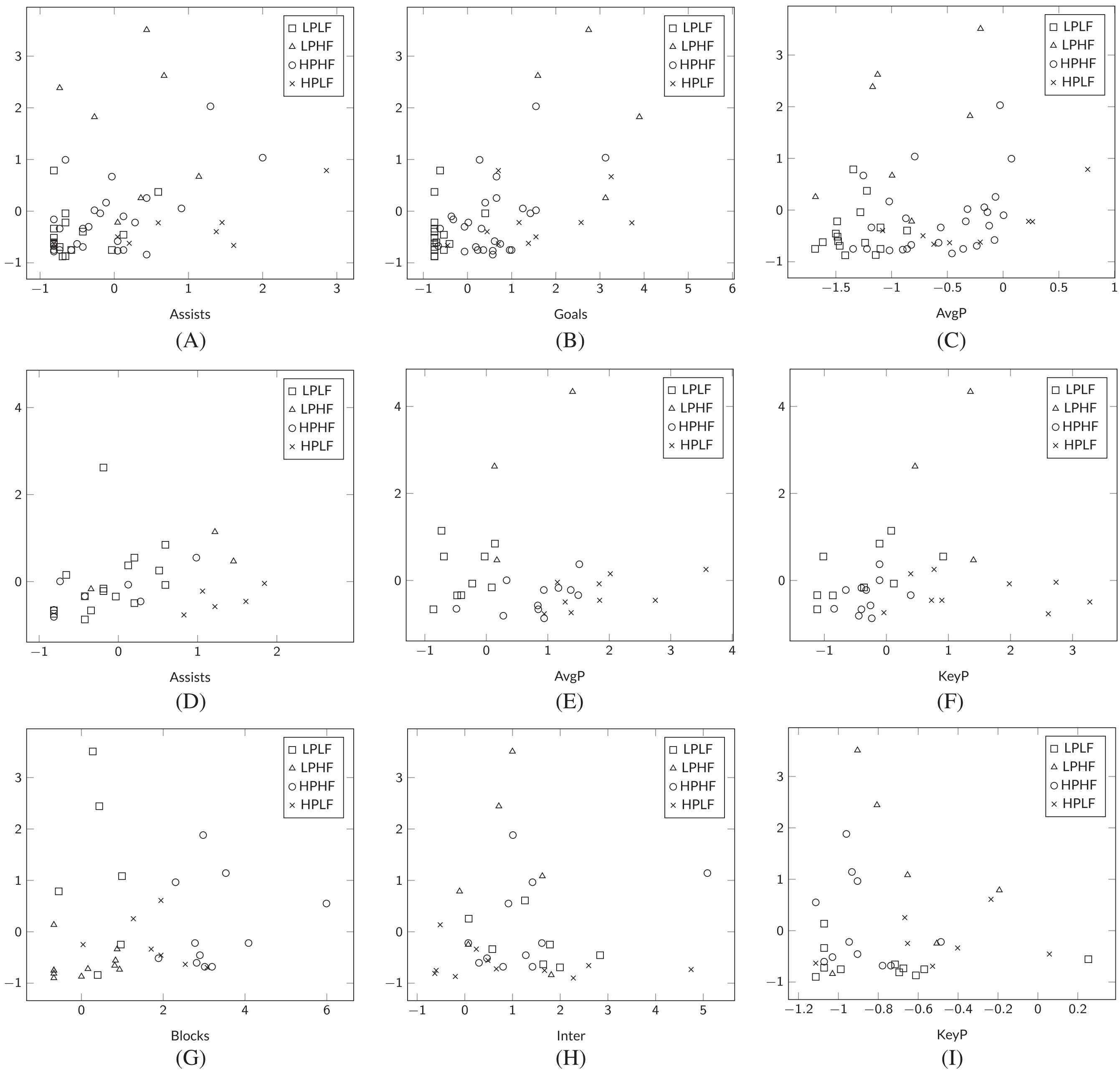

FIGURE 3 Clustering CF, CM, and CB by using Feature Selection (ylabel = *Fee*). A, Selected skills assists for CF; B, Selected skills goals for CF; C, Selected skills AvgP for CF; D, Selected skills assists of CM; E, Selected skills AvgP of CM; F, Selected skilss KeyP of CM; G, Selected skills blocks of CB; H, Selected skills inter of CB; I, Selected skills KeyP of CB

In this regard, Figure 3G to 3I can be interpreted differently than *CF* and *CM*, because the lower the ranking of football clubs, the lower the transfer fee of football players. Furthermore, since tactics use almost defensive counterattack tactics, the value of defensive skills is high.

5 | CONCLUSION

5.1 | Discussion

In this paper, data analytics is closely related to sports field and start to be concerned about how to contribute. Especially, since the football transfer market is growing rapidly, we have been focusing on the transfer market. The existing transfer system of the football transfer market is being transferred due to the intuition of the board member and the coach. For instance, Alvaro Morata, mentioned in Section 1, is a representative player of existing transfer systems. He has only 15 goals in the EPL since the transfer, compared to a high transfer fee.

In this regard, we thought that when clubs were considering player transfers, they would be able to make a successful transfer based on a reasonable basis if they combined their intuition and the results of the data analytics. To analyze the data, we collected data based on the history of *Whoscored* and performed regression analysis to predict the performance of each player after transfer. Each position has different important variables, so a feature selection was conducted for each position. Finally, we clustered using only selected variables for each position.

As a result in the experiment, *HPLF* label includes the players with high performance of various variables as comparing with transfer fee. For instance, in 2015, *Diego Costa*, a player who was transferred to Chelsea FC with the low transfer fee comparing with his performance in that season (Chelsea FC won EPL title). However, clubs do not have to buy only the players who belong to *HPLF* label, because the pursued policies of each clubs are different. For example, lower-ranked clubs may be aimed not to be demoted by bringing in relatively low transfer fee players, while top-ranked clubs may be aimed winning of the league even with the highest transfer fee players. Hence, we expected that, based on the results in our approach, club's management and board member could have better results.

5.2 | Future work

There are several limitations in this paper. First, the transfer fee was decided not only as a player's performance, but also as a player's popularity, transfer maket's Inflation and Homegrown Player Rule, and so on. However, this paper mainly focuses on performance analysis. The second limitation is lacking of data. There are not many players whose transfer fee was subject to data disclosure. In addition, we wanted to collect data for three years before transfer, but some players were unable to collect data. *Whoscored* only collects records for major leagues in Europe, so there is no data or the players, which are too young and do not debut during three years ago. As a result, we think that the more reliable analysis would have been possible if there have more data. Furthermore, if there is more than one club who competes for one player, the transfer fee would continue to rise. By inserting an existing player's performance into the cluster, it is possible to estimate the approximate transfer fee for the cluster, even if not the exact transfer fee. However, if many clubs compete for one player, transfer fee will continue to rise, which is not taken into account in this study.

In this regard, for the future work, we take into consideration on the optimization of club-specific transfers that complement these limitations. Moreover, the extension of this study is expected to contribute not only football but also the sports market in various ways.

ACKNOWLEDGMENTS

This study was supported by the Chung-Ang University Research Scholarship Grants in 2019. In addition, this work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2017R1A2B4010774).

ORCID

Jason J. Jung https://orcid.org/0000-0003-0050-7445

REFERENCES

1. Rein R, Memmert D. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*. 2016;5(1):1410. https://doi.org/10.1186/s40064-016-3108-2
2. Valerdi R. Why software is like baseball. *IEEE Software*. 2017;34(5):7-9. https://doi.org/10.1109/MS.2017.3571583
3. Park EM, Seo JH, Ko MH. The effects of leadership by types of soccer instruction on big data analysis. *Cluster Computing*. 2016;19(3):1647-1658.
4. Yu Y, Wang X. World Cup 2014 in the Twitter World: A big data analysis of sentiments in US sports fans' tweets. *Comput Hum Behav*. 2015;48:392-400. https://doi.org/10.1016/j.chb.2015.01.075
5. Piccialli F, Jung JE. Understanding customer experience diffusion on social networking services by big data analytics. *Mob Netw Appl*. 2017;22(4):605-612.
6. Nguyen HL, Jung JE. SocioScope: A framework for understanding internet of social knowledge. *Future Gener Comput Syst*. 2018;83:358-365.

7. Min B, Kim J, Choe C, Eom H, Mckay R. A compound framework for sports results prediction: a football case study. *Knowl-Based Syst*. 2008;21(7):551-562. https://doi.org/10.1016/j.knosys.2008.03.016
8. Rohde M, Breuer C. Europe's elite football: financial growth, sporting success, transfer investment, and private majority investors. *Int J Financial Stud*. 2016;4(2):12.
9. He M, Cachucho R, Knobbe A. Football player's performance and market value. In: Proceedings of the 2nd Workshop of Sports Analytics, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD); 2015; Porto, Portugal.
10. Ribeiro A, Lima F. Portuguese football league efficiency and players' wages. *Appl Econ Lett*. 2012;19(6):599-602. https://doi.org/10.1080/13504851.2011.591719
11. Pearson G. Sporting justifications under EU free movement and competition law: the case of the football 'transfer system'. *Eur Law J*. 2015;21(2):220-238. https://doi.org/10.1111/eulj.12110
12. Reeder A, Adams I. What if… Abramovich bought the Spurs? *The Sporting Image: What If?* Preston, UK: SSTO Publications; 2010:241-248.
13. Kurakin A, Vitusevich S, Petrychuk M, et al. Influence of overheating effect on transport properties of AlGaN/GaN heterostructures. *Mater Res Soc Symp Proc*. 2006;955:119-123.
14. Chianese A, Marulli F, Piccialli F, Benedusi P, Jung JE. An associative engines based approach supporting collaborative analytics in the internet of cultural things. *Future Gener Comput Syst*. 2017;66:187-198. https://doi.org/10.1016/j.future.2016.04.015
15. Bui KHN, Jung JJ. Computational negotiation-based edge analytics for smart objects. *Information Sciences*. 2018;480:222-236. https://doi.org/10.1016/j.ins.2018.12.046
16. Yin S, Kaynak O. Big data for modern industry: challenges and trends [point of view]. *Proc IEEE*. 2015;103(2):143-146. https://doi.org/10.1109/JPROC.2015.2388958
17. Lee O, Jung JJ. Modeling affective character network for story analytics. *Future Gen Comput Syst*. 2018;92:458-478. https://doi.org/10.1016/j.future.2018.01.030
18. Beneventano P, Berger P, Weinberg B. Predicting run production and run prevention in baseball: the impact of Sabermetrics. *Int J Bus Humanit Technol*. 2012;2(4):67-75.
19. Bojanova I. IT enhances football at World Cup 2014. *IT Professional*. 2014;16(4):12-17. https://doi.org/10.1109/MITP.2014.54
20. Stefani R, Pollard R. Football rating systems for top-level competition: a critical survey. *J Quant Anal Sports*. 2007;3(3):1-22. https://doi.org/10.2202/1559-0410.1071
21. Hvattum L, Arntzen H. Using ELO ratings for match result prediction in association football. *Int J Forecast*. 2010;26(3):460-470. https://doi.org/10.1016/j.ijforecast.2009.10.002
22. Cuomo S, Michele PD, Piccialli F, Galletti A, Jung JE. IoT-based collaborative reputation system for associating visitors and artworks in a cultural scenario. *Expert Syst Appl*. 2017;79:101-111. https://doi.org/10.1016/j.eswa.2017.02.034
23. Hong M, Jung JJ. Multi-sided recommendation based on social tensor factorization. *Information Sciences*. 2018;447:140-156. https://doi.org/10.1016/j.ins.2018.03.019
24. Sæbø O, Hvattum L. Evaluating the efficiency of the association football transfer market using regression based player ratings. 2015.
25. Vishnuvarthanan G, Rajasekaran MP, Murugan PR, Subbaraj P, Vishnuvarthanan A. An unsupervised learning method with a clustering approach for tumor identification and tissue segmentation in magnetic resonance brain images. *Appl Soft Comput*. 2016;38:190-212. https://doi.org/10.1016/j.asoc.2015.09.016
26. Deng L, Yu D. Deep learning: methods and applications. *Found Trends® Signal Process*. 2014;7(3-4):197-387. https://doi.org/10.1561/2000000039
27. Mooi E, Sarstedt M, Mooi-Reci I. Cluster analysis. *A Concise Guide to Market Research*. Berlin, Germany: Springer; 2019:313-366.
28. Dhanachandra N, Manglem K, Chanu Y. Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. *Procedia Comput Sci*. 2015;54:764-771. https://doi.org/10.1016/j.procs.2015.06.090
29. Kohavi R, John G. Wrappers for feature subset selection. *Artificial Intelligence*. 1997;97(1-2):273-324. https://doi.org/10.1016/S0004-3702(97)00043-X
30. Kursa M, Rudnicki W. Feature selection with the Boruta package. *J Stat Softw*. 2010;36(11):1-13.
31. Charrad M, Ghazzali N, Boiteau V, Niknafs A, Charrad M. Stackelberg game for utility-based cooperative cognitive radio network. *J Stat Softw*. 2009;61:1-36. https://doi.org/10.1145/1530748.1530753

How to cite this article: Kim Y, Bui K-HN, Jung JJ. Data-driven exploratory approach on player valuation in football transfer market. *Concurrency Computat Pract Exper*. 2021;33:e5353. https://doi.org/10.1002/cpe.5353