A career in football: what is behind an outstanding market value?

Balazs Acs and Laszlo Toka

MTA-BME Information Systems Research Group
Faculty of Electrical Engineering and Informatics
Budapest University of Technology and Economics

Abstract. Identifying professional career path patterns is an important topic in sports analytics. It helps teams and coaches make the best transfers and team compositions. It also helps players find out what skills and how they need to improve to achieve their career goals. In this paper, we seek the player characteristics that mostly affect a player's evaluation. To this end, we first created three-year-long career path segments from the time series data of 4204 players, then we created clusters from each segment based on the market value change over the examined period. After the clustering we searched for professional career path patterns where the market value growth was outstanding. Then we identified the 5 most important features with dynamic time warping and calculated how these should change over the years to achieve this career path. Finally we validated our findings with binary classification. We found that it is possible to explain real life professional career path patterns based on outstanding market value growth with the information collected from the FIFA video game series data collection. We managed to evaluate the extent of how these characteristics should change over the years to achieve the desired career.

Keywords: football · soccer · professional career path· pattern search · time-series clustering · dynamic time warping · feature importance · binary classification.

1 Introduction

In the early 2000s sports analysts had a hard time, because of the limited data available. Fortunately, with the growing importance of the statistical approach to achieve competitive advantage, the quantity and quality of sport-related data has also increased. First baseball, NBA, NHL and NFL enjoyed the benefits of the analytical approach, but now with the explosive growth of available soccer data, there are plenty of areas where a solution or tool can be developed, which may lead to a competitive advantage for both teams and players. There are many sports analytics models ranging from the effects of situational variables on performance [1], through the importance of game context [2], to creating new measurement metrics like EPV [3].

An important topic in the field of football analysis is to determine which players have potential to achieve an extraordinary professional career. This helps both coaches and teams to identify the most promising players and to create a powerful team composition. In addition to all this, players can also benefit from this. They can adapt to the career schemes that are potentially good match to their capabilities. Our goal in this work is to find patterns in football player career paths in terms of outstanding market value growth and to tell which skill features determine success and how those can explain market value.

In Section 2 we present our data collection from Sofifa.com [4] and Transfermarkt.com [5]: we highlight the differences in the data from these two sources and the football market inflation observed therein. Next, in Section 3 we introduce the steps of modeling: the career path segmentation, clustering and the search for professional career path patterns where the market value growth was outstanding. In Section 4 we show the most important skill features of the outstanding players with the help of dynamic time warping and their dynamics during the examined years; afterwards a binary classification with LGBM is presented to validate the role of these features in practice. In Section 5 we discuss related work, and finally in Section 6 we conclude this work.

2 Player evaluation data: collection and preparation

Our goal was to collect data diversified by nationality, club, league, or international popularity. We used two sources for this type of data: Sofifa.com and Transfermarkt.com. From Sofifa we were able to collect 15 years of data about players included in the FIFA database. The data from this site plays a big role in this analysis, because it contains 21 different skill scores (e.g., dribbling, short-passing, finishing), market values, wages and other personal information (e.g., age, weight) for each player. All skill variables are stored in a range from 1 to 99, the higher is the better. It is important to note that the different positions (e.g., ST, GK, CAM) require different skills. For example, for a goalkeeper the gk_handling skill is more important than finishing. Besides the skill and personal features, the market value is a key element.

Collecting the data from Sofifa was not enough because of two reasons. First, between 2007 and 2011, the market value of players is missing from the Sofifa database. As market values are essential in our analysis, we had to collect them from another site, i.e., Transfermarkt.com. Furthermore, within the Sofifa player pricing data there is a significant difference in the year-over-year change of the mean market values compared to the Transfermarkt values, and there is a great fluctuation in the values between 2012 and 2016, raising suspicion about the quality of data.

The range of the collected data is 15 years (from 2007 to 2021) from both sources. In order to deal with the distance between the high-end and low player values, we performed a logarithmic scaling transformation of prices. We normalized all market values by dividing them with the highest value so the value range during the modeling was between 0 and 1. Moreover, we transformed the

age feature by subtracting 16, so the feature shows rather the time spent in professional career, being between 0 and 20 in our dataset.

2.1 Market value differences between Transfermarkt and Sofifa

We found a significant difference between the player values we collected from the two sources. First of all we wanted to figure out if the two datasets are from the same population. For this purpose we used the Mann-Whitney U test and Kruskal-Wallis H test. We tested the market values every year, from 2012 to 2021 (before 2012 the Sofifa dataset had missing pricing values). We denote the Transfermarkt dataset as TR and the Sofifa dataset as SO throughout this paper. Except for 2012, the TR and SO datasets were different, we had to reject H0 at 0.05 significance level: we can clearly state that the market values reflected in TR and in SO are different.

Moreover, the mean year-over-year changes are also significantly different: in Figure 1 it is clear that the price evolution is far from being the same in the two datasets. Since the market values before 2012 were not available in the SO dataset and the fluctuation has been much greater over the years than in TR, we decided to use only the TR prices during the analysis.

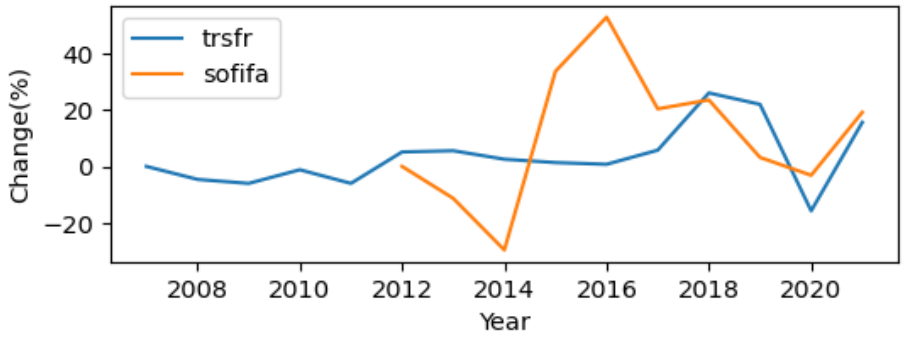

Fig. 1. Transfermarkt vs. Sofifa mean year-over-year market value change

2.2 The 2014-16 market value boom

In Figure 1 it is noticeable that something happened between 2014 and 2016. The SO dataset shows a strong market value increment in this period, and the TR values are also increasing during and after this phenomenon. After the decrease from 2012 to 2014, in 2015 there was a 33% growth and from 2015 to 2016 this growth increased to 53% in SO. This exceptionally large jump between 2014 and 2016 is due in part to affluent Arab and Russian investors. In these years United Arab Emirates, Qatar and Saudi Arabia suddenly appeared in the list of

the top 20 spenders in the football market across the world. In 2014 Al Arabi (Qatar Stars League) spent over 50 million dollars on the transfer market and with this Al Arabi was the eighth placed club inside the top 10 spenders. In the 2016 transfer season, Arab football teams spent over 200 million dollars on transfers [6].

2.3 Handling the football market value inflation

We also tackled the issue of price inflation. The inflation in the world of football seems to supersede the regular monetary inflation. For example, 1 British pound in 1990 was worth 2.27 in 2019, but in football 1 pound of 1990 would be worth about 40 now. As a stellar illustration: while in 1989 the whole squad of Manchester United was worth 20 million pounds [7], today the most valuable player of the world is Kylian Mbappé with 144 million pounds. With this observation we calculated the inflation rate of market values over the collected 15 years. We tried different approaches to handle inflation: we adjusted the market values of each year to match the mean, median and the third quartile statistics with those of the latest year values. By doing so our intention was to set the values of each year to their present value as close as possible.

We found that the best statistics was the third quartile (Q3): we set the Q3 in 2021 as the base value, being 1.8M Euros, and linearly scaled each year's values to bring their Q3 to this value. After this transformation every year had 1.8M Euros as Q3: the YOY became uniform, the sudden changes in pricing and even the effects of the 2014-2016 market value boom disappeared.

3 Time series analysis of player value

In this section we present the steps we made to successfully create our model that identifies if a player had an outstanding market value growth during the examined years of his career. First, we created career path segments for each player based on their professional career, then we created clusters from these segments. Based on the clusters we could find the patterns of the different player career paths. In the end, we find the most important features that affect the outcome of the professional careers of players.

3.1 Career segmentation

Before the time series clustering, we made some changes in our dataset. For a robust time series analysis we chose players with more than 9 years of data, and removed the rest. Our aim was to create clusters on the dynamics in each 3-year-long career path segment. For example, if a player had 9 years of data, then we were able to create 3 segments with length of 3 years, and we got 3 cluster labels for the career segments of that player. We removed goalkeepers from our dataset, because their most important skills are exceptional. For example, comparing

gk_handling or gk_diving to skills in the majority, such as finishing, dribbling, marking and so on, could have been a problem in our evaluation.

Before this filtering step the dataset had 43580 individual players. After the removal, our dataset was reduced by about 90%.In the end we created 3-year-long career path segments for each player with 9, 12 and 15 years of professional career time from a total of 4204 players.

3.2 Player clustering based on value dynamics

For clustering the 3-year-long segments, we used K-means. To find the best K value we used both the elbow method and silhouette score. With both methods 3 was the optimal number of clusters. Figure 2 shows the 3 clusters we created from the first career path segment of players with 9 years of professional career time. On the X axis we can see the years passed (0,1 and 2) and Y axis shows the current market value transformed by standard scaler. All the individual segments are indicated with black lines and the red line is the barycenter.

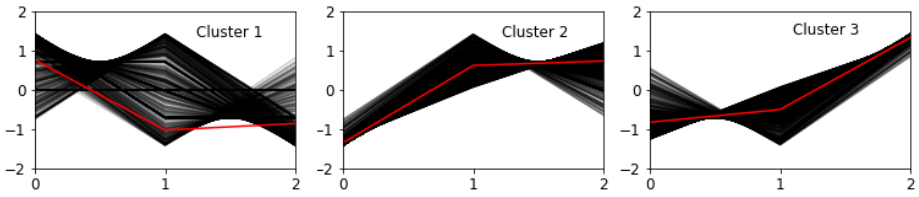

Fig. 2. First segment clusters with K-means

We labeled the clusters based on the player's path of development in market value: decreasing (1), stagnating (2) and increasing (3). Cluster 3 in Figure 2 shows strong increasing trend; in cluster 2 the player market value grows in the first year, then it stagnates; in cluster 1 we can see a strong decreasing behavior.

We performed the same clustering on the remaining career segments, and arrived at similar results: in all cases the optimal number of clusters was 3, and the same increasing, stagnating and decreasing trend were grouped in the clusters, the only difference was in the intensity of decline and growth of the market values. Therefore we apply the same labels, i.e., 1, 2 and 3, in every career path segment of each professional career time series (9, 12 and 15 years).

3.3 Pattern search

With the help of the cluster labels we created we were able to find patterns in the whole career paths. Our aim was to find players with extraordinary market value growth in the examined years, we call them as "selected players" from now on. To this end we looked at the cluster labels of the first 3 segments of each player (as an aging player might end up in worse clusters at the end of their

career), and selected the player if the sum of cluster labels was exactly 9, i.e., all increasing segments. By selecting the players with the highest possible sum of cluster labels we wanted to find players whose market value never stagnated or decreased significantly during the examined years. In Figure 3 we can see the value evolution of two players who were classified as "selected players". X axis shows here the professional career time of the player in years (between 0 and 20), and on the Y axis we can see the normalized market value of the players (the log base 10 form of market value was divided by the highest market value, it is between 0 and 1).

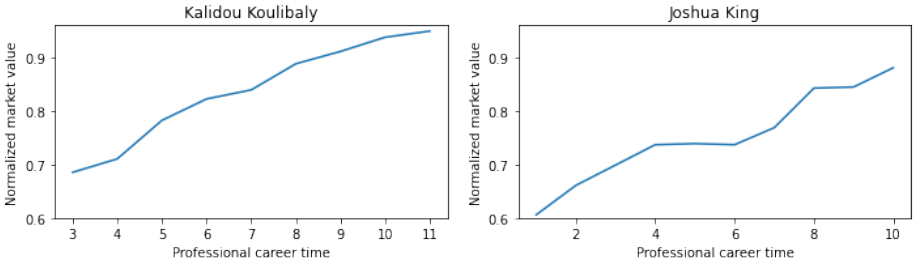

Fig. 3. Players with outstanding market value growth during the examined years

4 Finding outstanding players based on market value change

We want to find out which skill features play the biggest role in becoming an extraordinary player. In addition, we want to know exactly how these variables should evolve over one's career. In this section we present the model we built to determine whether a player is within the selected ones or not with the use of the previously mentioned skill variables.

We created the list of selected players with the pattern search method discussed in the previous section. We found 138 players who have met the conditions, and created the label values accordingly.

The next step was to find the most important features. We searched for the largest distances between the selected and other players features: we used dynamic time warping (dtw) for this purpose. First we calculated the distance of features between the selected players and we got 18 906 distance records for each feature. Next, we did the same on the rest of the players. In order to reduce the computation time, we only used every 10th player for the latter. In the end we got 165 242 distance records for each feature from the other players. Finally we compared the distances with the scores given by dtw. Figure 4 shows the final feature distances between the selected and other players. Each skill variable had

a distance score between 25 and 13. X axis shows the feature ID and Y axis shows the distance scores.

For the most important features we decided to select the skills with a distance score above 20, because a larger gap was observed between the distances above and under 20. The highest feature distance score under 20 was 18 and most of the remaining features were in the interval of 15 and 17. Therefore the 5 most important features are: Marking (24.296), which is the ability of a player to prevent the opposing player from obtaining the ball; Standing_tackle (23.871), the ability to perform a tackle to capture the ball; FK_accuracy (20.282), measures the free kick accuracy and the chance of scoring a goal from it; Finishing (20.264), which determines the ability of a player to score from an opportunity; and Long_shots (20.180), the accuracy of shots from long distances. After we

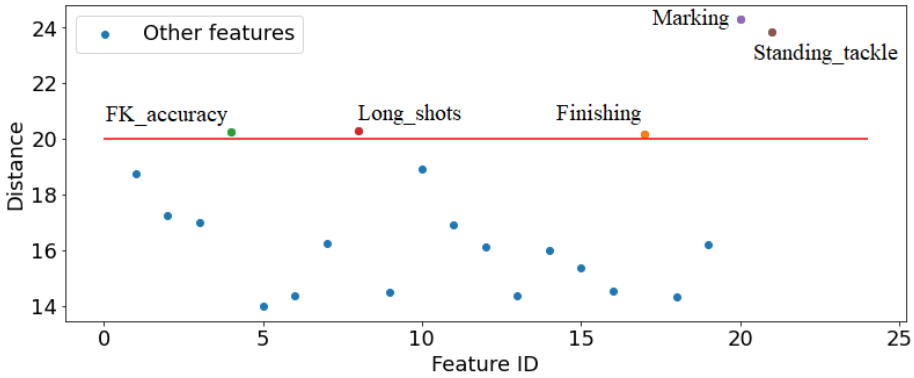

Fig. 4. DTW distances

found the top 5 features, we wanted to know exactly how they changed for the selected and other players during the period under review. We calculated the mean change of every skill and compared them. In Figure 5 we can see the results. Each dot represents the mean change of the skills over the 9 years. The red number shows the mean difference of the skill development between the selected players (blue) and other players (orange). The ranking is almost the same with so much difference that the finish came in second place in terms of difference. For example to become a selected player, the player must boost the Marking skill by +7.62 compared to the other players, or improve Finishing by +6.45 over 9 years.

To check our findings in practice, we created a binary classification. In this classification our predictors were the top 5 features we selected and we wanted to predict if a player is selected or not. We used an LGBM classifier for this task and we measured our model with AUC. We used a train-test split in order to train our model and validate the findings on the test set. The model AUC score with the top 5 features was about 0.71. If we added the remaining features the

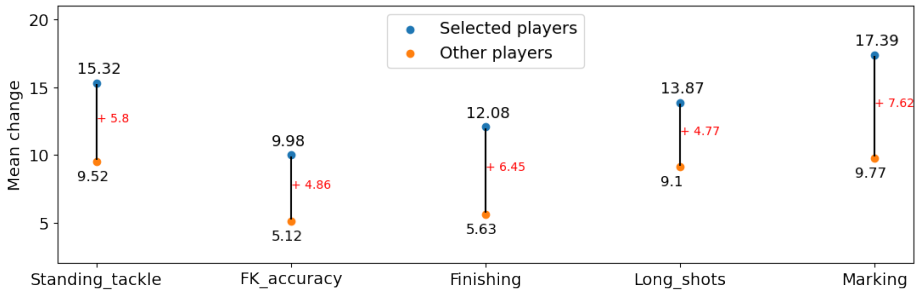

Fig. 5. Difference between the mean improvement of the top 5 features of the selected and the other players

score improved to about 0.75, but we can state that the 5 selected features played the biggest role during the classification. This relatively low score is highly due to the lack of available data. Sofifa only stores player data from 2007 and we could only use the characteristics of the 4204 players (138 selected) whose data were available for a time interval with the required length.

To check if we can achieve different results by augmenting the dataset, we created career segments with an overlapping method. Instead of creating the segments from 1-3, 4-6 and 7-9 years of professional career, we created them from 1-3, 2-4, 3-5 years and so on. With this method the number of observations has increased significantly. In total 12166 player career segments were examined: 3148 selected and 9018 other players. Therefore the ratio of 1:400 (selected:other) changed to 1:3.

In addition we considered to change our approach of clustering. Previously we created 3 clusters based on the elbow method and silhouette score. In Figure 2,

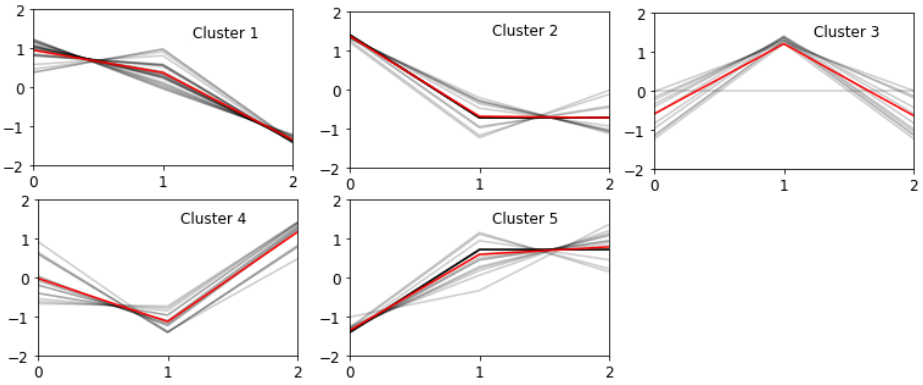

Fig. 6. Modified clusters

it is observable that the barycenters are correct but some players might have been misclassified. To make sure every item is correctly clustered we created 5 clusters; in Figure 6 we can see almost the same patterns as before and we labeled them as usual. Clusters 1 and 2 are decreasing (1), Cluster 3 is stagnating (2) and finally, Clusters 4 and 5 are increasing (3) trends.

After the clustering we did the same steps as before. Surprisingly the dynamic time warping results were exactly the same. The same 5 features had outstanding impact on careers, as we showed in Figure 5. After a train-test split and binary classification with LGBM, the AUC score was about 0.70 which is almost the same as the results from the previous model. This exercise validates the robustness of the model, and we can clearly state that even if we increase the number of observations the results will not change drastically.

5 Related work

Many related works have appeared in this topic. Here we discuss what are the key differences compared to our work, and what conclusions can be drawn.

Bettina Schroepf and Martin Lames searched for career patterns in German football youth national teams [8]. They managed to find 8 typical career types. It has been found that the careers of youth players last up to 1 or 2 years in Germany and only a few players can achieve long-lasting career. It is interesting how big is the churn rate by young players in major European football nations, so it is already a privilege here for someone to be among the pros. In our paper we have broaden the search and tried to find career paths of adult players around the globe. A recent study revealed that in Portugal the length of a career as a football player is decreased, but the years of youth career increased: it follows that the career path started earlier in the last 3 decades [9]. Remaining in Portugal, Monteiro et al. identified that the best result the players performed was in the age of 27 and they ended their career at around 33 [10]. Other researchers revealed that the peak performance by female football players is around the age of 25 [11] or between 25 and 27 [12] in case of male players. In our study we preserved the players between the age of 16 and 34 and the majority of the examined players were between the age of 22 and 27.

Identifying the career potential of athletes is a very difficult task. Coaches play a big role in this aspect, because they have the closest relationship with the players. However they also encounter difficulties. In the article of Cripps, A. J. et al., the authors found that coaches can predict the career outcome more accurately for late maturing athletes, but they are less accurate for early maturing players [13]. It is important to keep in mind that the way of maturation can easily affect the career path both positively and negatively.

It is also interesting, how it is possible to draw a parallel between psychology and performance in terms of success. Schmid, M. J. et al., found patterns in rowing by connecting these variables. They measured the proactivity, ambition and commitment of the athletes for the past year and 30 months later. As a result, if a highly motivated rower had poor performance in the past year, it was

more likely that he/she performed at a very high level in the future. Athletes with low moral and motivation were more likely to drop out or perform weaker [14]. Good motivation is a key to perform better. Highly achievement oriented players have a better chance to accomplish an outstanding career [15]. We only used skill/performance variables but based on these statements, there is potential in the introduction of certain psychological features into our model.

Vroonen, R. et al. [16] predicted the potential score (available on Sofifa.com) of professional football players. They selected a player and searched for similar players from the same age. Based on the evolution of the latter, they predicted if he/she had great potential score in the future. In contrast, in our paper we recognized continuous market value growth patterns and we have shown which skill development is required to achieve this goal. We did not predict the potential scores of Sofifa, rather we paralleled the available information from Sofifa with the development of real market prices and we successfully explained the real life career path patterns we were looking for with them.

6 Conclusion

In this paper we presented an approach to identify patterns in football player career paths based on extraordinary market value changes and skill features that are responsible for excellence. The time segmenting and time series clustering method we applied was successful and we managed to find the patterns we were looking for. Next we found the 5 most important features and the exact dynamics of how these variables should evolve to enter the high class of football players. The need to develop Marking (+7.62) and Finishing (+6.45) skills has been outstanding over the years, but Standing_tackle (+5.8), Free-kick_accuracy (+4.68) and Long_shots (+4.77) were also strongly needed to achieve an outstanding career path. Finally we validated the results in practice with binary classification, the AUC score of the model with the 5 selected features was 0.71 and with the other skill features was 0.75. We proved that despite the fact that at first we worked with little data, the model is robust, because with increased amount of data we got the same score results. As the increased sample size could not affect the results, there is a need for additional features to enhance the model. In the future there is a potential to improve these findings by involving psychological or situational variables.

Teams, coaches and players can have a wide range of benefits from this study. Teams and coaches can use the lessons learned for strategical decisions, for example how to train youth players in order to sell them early with high return. This could bring a profitable decisions and additional source of income, vital for today's competitive environment in the top leagues. It is also a good indicator for coaches to identify future high class players. In the future we want to broaden the scope of the model so we can apply the findings on career goals other than high valuation, like playing in Champions League finals, being a member of the national team, transferring to a higher division or ending up contracted to a desired club.

Acknowledgment

Project no. 128233 has been implemented with the support provided by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the FK_18 funding scheme.

References

1. Gómez, M. Á., Lago-Peñas, C., Pollard, R.: Situational Variables. In: McGarry, T., O'Donoghue, P., Sampaio, J. (eds.) Routledge Handbook Of Sports Performance Analysis, pp. 277–287. Routledge, London (2013). https://doi.org/10.4324/9780203806913
2. Power, P., Ruiz, H., Wei, X., Lucey, P.: Not All Passes Are Created Equal: Objectively Measuring the Risk and Reward of Passes in Soccer from Tracking Data. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1605–1613. Association for Computing Machinery, Halifax, NS, Canada (2017)
3. Fernández, J., Bornn, L., Cervone, D.: Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer. In: MIT Sloan Sports Analytics Conference 2019, Boston (2019)
4. Sofifa, https://sofifa.com/. Last accessed 21 June 2021
5. Transfermarkt, https://www.transfermarkt.com/. Last accessed 21 June 2021
6. Raisi, O. A.: The Economics Of Middle East's Football Transfers, https://www.sportsjournal.ae/the-economics-of-middle-easts-football-transfers/. Last accessed 21 June 2021
7. Christou, L.: The true extent of spiralling inflation in football's transfer market, https://www.verdict.co.uk/football-transfer-market-inflation/. Last accessed 21 June 2021
8. Schroepf, B., Lames, M.: Career patterns in German football youth national teams – A longitudinal study. International Journal of Sports Science & Coaching **13**(3), 405–414 (2018)
9. Carapinheira, A. et al.: Career Termination of Portuguese Elite Football Players: Comparison between the Last Three Decades. Sports **6**(4), 155 (2018)
10. Monteiro, R. et al.: Identification of key career indicators in Portuguese football players. International Journal of Sports Science & Coaching **15**(4), 533–541 (2020)
11. Barreira, J.: Age of Peak Performance of Elite Women's Soccer Players. International Journal of Sports Science **6**(3), 121–124 (2016)
12. Dendir, S.: When Do Soccer Players Peak? A Note. Journal of Sports Analytics **2**(2), 89–105 (2016)
13. Cripps, A. J., Hopper, L. S., Joyce, C.: Can coaches predict long-term career attainment outcomes in adolescent athletes?. International Journal of Sports Science & Coaching **14**(3), 324–328 (2019)
14. Schmid, M. J., Conzelmann, A., Zuber, C.: Patterns of achievement-motivated behavior and performance as predictors for future success in rowing: A person-oriented study. International Journal of Sports Science & Coaching **16**(1), 101–109 (2021)
15. Zuber, C., Zibung, M., Conzelmann, A.: Motivational patterns as an instrument for predicting success in promising young football players. International Journal of Sports Science **33**(2), 160–168 (2015)

16. Vroonen, R. et al.: Predicting the potential of professional soccer players. In: Davis, J., Kaytoue, M., Zimmermann, A. (eds.) ECML PKDD 2017, Proceedings of the 4th Workshop on Machine Learning and Data Mining for Sports Analytics, vol. 1971, pp. 1–10. Springer, Skopje (2017)