PREDICTIVE MODELS APPLIED IN SPORTS MANAGEMENT – LITERATURE REVIEW ON RESEARCH TRENDS

https://doi.org/10.47743/jopafl-2022-23-11

Irina-Cristina COJOCARIU
Alexandru Ioan Cuza University of Iaşi,
Iaşi, Romania
cojocariu.irina96@yahoo.com

Abstract: *In recent years, numerous studies have been conducted to obtain by extracting the most accurate estimate of the main parameters in a given field. The techniques were diverse, and the main purpose was to identify how information can become useful knowledge. My area of interest, sports prediction, is constantly evolving, so many organizations have begun to focus on these methods that can provide them with valuable data. Therefore, this article is actually a literature review on how sports data is exploited. On this basis, I can present an overview of what has been studied, research proposals, topics addressed, algorithms and technologies used and future opportunities. Analysing these discoveries I want to offer a mining potential in this field and to attract as many researchers as possible to research the subject of sports predictions.*
Keywords: *Sports analysis, prediction models, prediction of ticket sales, match results*

Introduction

The concept of prediction captures a new facet of the digital age that facilitates revenue growth in various fields. Sports prediction is usually treated as a classification problem, with only one class (win, lose or draw) to be predicted Prasetio and Harlili (2016). The use of a structured experimental approach to the problem of predicting sports results is useful for obtaining the best possible results with a data set. Predictive models can also be used to build data products, for this purpose there is a system of recommendations that could help clubs make different decisions. Scientific research has not only tried to identify a model applicable to a sports club. After identifying the framework, the researchers tried to generalize the model so that it could be used by as many clubs in different countries of the world. In this article I want to understand how the prediction of results is thought and the generalization of predictors so that they can be used on a large scale. Moreover, I will identify which are the main ones directions regarding sports analysis and what are the models, respectively the variables used in the analysis.

Research background

The beginnings of sports analysis focused mainly on the analysis that referred to the proceeds from the sale of tickets to matches. In this regard, I identified a two-way specialized studies: event management (which includes analysis of ticket sales, participation of fans in stadium matches) and sports performance of players. The first category can be divided into two other directions as follows: the analysis of tickets sales at matches and the presence of fans at the stadium. The presence at the stadium is a major source of income for all sports teams, theoretical and empirical research on the demand for participation has been an integral part of the sports economy. The two oldest empirical

studies of determinants were made in the 1970s by Noll (1974) and Demmert (1973). Each study carefully explained a variety of factors that could change demand, including control variables for local income, the age of the stadium, the substitutes of availability, the success of the franchise and the population of the local market.

Many studies have been conducted on participation in matches. It will be noticed that most of them are econometric studies whose objective is to highlight which are the factors determinants of demand. As for the variables analyzed in order to obtain as high an accuracy as possible, the number of tickets that would be sold for a match, the research shows that the number of points scored by home and away team in the previous five matches is taken as a significant factor for participating in matches. And in 2007, in another article in the domain are investigating the impact of big players that are proving to have an impact on increasing match attendance (Brandes et al., 2007). As we could observe at the level of studies there are two types: controllable variable (opponent, match day, ticket price) and variables that cannot be controlled (weather, atmospheric pressure, the wind). And, for this reason, it will be interesting to obtain information about these variables and whether there are really significant connections with total receipts from ticket sales to matches.

Since the pandemic affected the participation of the fans in the matches played at the stadium, I turned to another direction that can be analyzed - the performance of the players. Another less analyzed direction is the prevention of injuries. This side requires a history of players which is difficult to centralize without digitization. Rossi et al., (2018), using variables such as position, age, height, weight, gps coordinates, but also variables related to distance, speed, number of previous injuries and previously played matches analyzed the predisposition of 26 players in depending on the variables mentioned above. On the other hand, another study (Bongiovanni et al., 2020) using variables such as anthropometric features corrected arm muscle area, arm muscle circumference, right and left suprapatellar girths applied for the analysis of a football academy in Italy focused on physical performance prediction. Also in this direction, I included another research (Dijkhuis et al., 2021) to which we add the most replacements in the 50th minute, but also in the 60-90 interval, position, acceleration, energy, distance cover, distance in speed category , energy expenditure in power category.

Another facet of sports analysis is the accessibility of data that show the performance of football has facilitated recent advances in soccer analysis. The so- called football journals (Luke et al., 2018) which capture all the events that take place during a match, are one of the most common data formats and have been used to analyze many aspects of soccer, both to the team (Cintia et al., 2016) and individual levels (Cintia et al., 2015). Of all the open issues in soccer analysis, the data-based assessment of a player's performance quality is the most difficult, given the lack of ground truth for that performance assessment and a consistency in adding or retrieving this information.

Methodology

The aim of the paper is to take into account the specialized papers in the field studied, sports prediction, and for this reason I want to analyze the most relevant studies especially in the technological context - models used in predictive. Using the relevant keywords in the Google Scholar and web of science engines, I identified the main

clarifications on the subject, according to their relevance with the subject. I will present continuously the results obtained using predictive models and then the main variables studied.

Results

Results related to predictive models used

My focus is to identify papers from which I can extract information related to obtaining and analyzing predictive variables. If the results were used in the specialized studies or the data were trained to obtain predictions, and more importantly, where the data were extracted from, this being a rather big problem for research. In recent years, it has shown us that with the ever-increasing technology in our lives, more and more precise analyzes are needed. In this sense, sports analysis is interconnected with various camera devices, sensors, etc. One of the options for performing sports analysis is the video summary. Thus, small video slots can summarize the most important actions in a match. In addition to the frames needed to capture images, excitation event detection is also required (Zawbaa et al., 2012). In this sense, Bagadus was developed, a prototype that has a built-in sensor that aims to create video summaries by calibrating the cameras to form a panorama (Stensland et al., 2014). Another approach refers to detecting the key points of a match related to different parameters such as: correspondence with the ball, time dependent, not directly dependent on the ball, etc. or on a player movement system (Stein et al., 2017).

The introduction of tools and predictive models based on Machine Learning (ML) is another facet of the field. R is the language that covers data analysis, modeling and other operations based on statistical analysis, and according to a study by Kaggle 12% of respondents use this language for Data Science activities, being in the top 3 preferences (Data Science Survey, 2018). Two of the most popular classification and regression tree building techniques that are an integral part of Machine Learning are the Random Forest (RF) and Extreme Gradient Boosting (XGB) models (Breiman et al., 1984). For Random Forest the models are created based on the tidymodels framework (Kuhn et al., 2020), and for Extreme Gradient Boosting using the xgboost package (Chen et al., 2020). In a paper that focused on creating a predictive model built on the weather variables of the day of the match, the stage of the match, but also the calculation of the performances of the two participating teams in the last 4 matches played was obtained in addition to a satisfactory accuracy of the variables that influence the sale of tickets to soccer matches. Following the analysis performed using both Random Forest and XGB algorithms, it was observed that the latter is much more accurate, giving a higher score to the season and the environmental conditions of the match day (Fotache et al., 2021).

The main variables included in the studies

The variables of a research differ depending on the goal to be achieved and the area of interest. In the following I will present both variables related to previous matches and variables obtained during the current match used in predicting the results. Some research has attempted to determine the prediction of the winner and the loser respectively, based on possession of the ball and the analysis of approximately 20 actions during the match

(Capobianco et al., 2019). Other papers focused on the classification of the most important variables analyzed, which resulted in over 60% accuracy for predicting the results of a match using Random Forest model (Igoshkin, 2014). For those who want to introduce statistical analyzes, Berrar et al. (2019) present the Poisson and Bayesian models used in predicting outcomes. A probabilistic movement and zones of control are taken into account, thus dividing the field. Depending on where the ball was in the field, they could later establish control areas. In this way they analyzed 40 different events. An example of a challenge regarding sports prediction is the 2017 SoccerPrediction Challenge proposed by Kaggle in which 68 teams competed for soccer outcome prediction. Some of the variables considered by the participating teams are: season, league, date, home team, opposing team, number of goals for each team, goal difference, previous matches for competing teams (Dubitzky et al., 2019).

Regarding the obtaining of the data to be processed, two of the sites that offer free access are champinat.com which is easy to use to search for matches for each season and get information about form, focus and history by analyzing the match information page. Statoo.com is useful because it has a table for each time of the season, so the information is based on scores, positions, etc. can be exported in an easy way. Brooks et al. (2016) use the position and destination of a pass by analyzing passes, shots and tackles. Based on these, the individual strategy is determined, but also that of the team. The field is divided into 18 zones, and possession is determined by at least 3 passes between players of the same team. Because we follow the actions in a match, their number can be very high and their appearance low. For the Belgian Football Division, data were taken for 576 matches, of which approximately 100 actions were analyzed. In this case, VIF (variance inflation factor) was used to delete unimportant variables (Geurkink, 2021).

Conclusion and discussion

The field of sports prediction is an important economic and social factor of regional development worldwide. Sports innovation is an emerging field of research that links sport with management and good innovation practices. Innovation in sports is seen in new technologies, equipment, strategies and training improvements, and to have this overview we used a documentary study on a limited number of articles. Content analysis has as main characteristics objectivity, systematic character by creating explicit rules and their consistent application, as well as a quantitative character by which it is desired to count some occurrences. Supporting these characteristics, I can affirm that there is a trend in the area of predictive analysis focused on team performance, to the detriment of fan-focused analysis. The main focus is on how to play inside football clubs and less on how fans can be attracted to matches, as the variables on which the first criterion depends are easier to adjust. I also noticed that the number of variables differed from research to research, but in essence all those analyzed took into account the selective variables, according to the place and performance of the host team and the guest team. The limitations of the research are represented by the establishment of a number of articles analyzed and was based on a syntax that applied to the four databases from which the analyzed articles were extracted. This research gives me the opportunity to formulate a future research direction through which I want to make a generally valid prediction model of the results for soccer matches using a list of variables identified in the research analyzed previously.

References

1. Berrar, D., Lopes, P., Davis, J., & Dubitzky, W. (2019). Guest editorial: special issue on machine learning for soccer. Machine Learning, 108(1), 1- 7. https://doi.org/10.1007/s10994-018-5763-8
2. Bongiovanni, T., Trecroci, A., Cavaggioni, L., Rossi, A., Perri, E., Pasta, G., & Alberti, G. (2021). Importance of anthropometric features to predict physical performance in elite youth soccer: A machine learning approach. Research in Sports Medicine, 29(3), 213-224. https://doi.org/10.1080/15438627.2020.1809410
3. Brandes, L., Franck, E., & N esch, S. 7. Local Heroes and Superstars. Journal Of Sports Economics, 9(3), 266-286, https://doi.org/10.1177/1527002507302026
4. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
5. Breiman, L., Friedman J.H., Olshen R.A. and Stone, C.J, 1984.Classification and Regression Trees. Wadsworth & Brooks. Monterey, CA
6. Brooks, J., Kerr, M., & Guttag, J. (2016, August). Developing a data-driven player ranking in soccer using predictive model weights. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 49-55).
7. Bunker, R., & Thabtah, F. (2019). A machine learning framework for sport result prediction. Applied Computing And Informatics, 15(1), 27-33, https://doi.org/10.1016/j.aci.2017.09.005
8. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
9. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Li,Y. (2020). xgboost: Extreme gradient boosting. R package version 0.90. 0.1.
10. Cintia P., Giannotti F., Pappalardo L., Pedreschi D. and Malvaldi, M, 2015. The harsh rule of thegoals: Data-driven performance indicators for football teams. InProceedings of the IEEE International Conference onData Science and Advanced Analytics. https://doi.org/10.1109/DSAA.2015.7344823
11. Data Science Survey – 2018, https://www.kaggle.com/sudhirnl7/data- science-survey-2018
12. Dijkhuis, T. B., Kempe, M., & Lemmink, K. A. (2021). Early Prediction of Physical Performance in Elite Soccer Matches—A Machine Learning Approach to Support Substitutions. Entropy, 23(8), 952. https://doi.org/10.3390/e23080952
13. Dubitzky, W., Lopes, P., Davis, J., & Berrar, D. (2019). The open international soccer database for machine learning. Machine Learning, 108(1), 9-28. https://doi.org/10.1007/s10994-018-5726-0
14. Fotache, M., Cojocariu, I. C., & Bertea, A. (2021, June). High-Level Machine Learning Framework for Sports Events Ticket Sales Prediction. In International Conference on Computer Systems and Technologies' 21 (pp. 55-60)
15. Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics, 28(2), 337-407. https://doi.org/10.1214/aos/1016218223
16. Geurkink, Y., Boone, J., Verstockt, S., & Bourgois, J. G. (2021). Machine learning-based identification of the strongest predictive variables of winning and losing in Belgian professional soccer. Applied Sciences, 11(5), 2378. https://doi.org/10.3390/app11052378
17. Joel Brooks, Matthew Kerr, and John Guttag. 2016. Developing a data- driven player ranking in soccer using predictivemodel weights.
18. InProceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and DataMining. 49–55.
19. Kemper, C., & Breuer, C. (2016). Dynamic ticket pricing and the impact of time – an analysis of price paths of the English soccer club Derby County. European Sport Management Quarterly, 16(2), 233-253, https://doi.org/10.1080/16184742.2015.1129548
20. Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26, p. 13). New York: Springer.
21. Kumar, G. (2013). Machine learning for soccer analytics. University of Leuven.
22. Luke Bornn, Dan Cervone, and Javier Fernandez. 2018. Soccer analytics: Unravelling the complexity of the beautifulgame.Significance15, 3 (2018), 26–29.doi: https://doi.org/10.1111/j.1740-9713.2018.01146

23. Luo, G. (2016). A review of automatic selection methods for machine learning algorithms and hyper-parameter values. Network Modeling Analysis in Health Informatics and Bioinformatics, 5(1), 1-16. https://doi.org/10.1007/s13721-016-0125-6
24. Masoud, N. (2020). Moneyball: Sports Analytics in Soccer to Predict Performance and Outcomes. Experfy Insights, from https://www.experfy.com/blog/moneyball-some-insights-to-soccer- analytics
25. Mondello, M., & Kamke, C. (2014). The introduction and application of sports analytics in professional sport organizations. Journal of Applied Sport Management, 6(2), 11.
26. Morgulev, E., Azar, O. H., & Lidor, R. (2018). Sports analytics and the big- data era. International Journal of Data Science and Analytics, 5(4), 213- 222. https://doi.org/10.1007/s13721-016-0125-6
27. Patel, D., Shah, D., & Shah, M. (2020). The Intertwine of Brain and Body:A Quantitative Analysis on How Big Data Influences the System of Sports. Annals Of Data Science, 7(1), 1-16. https://doi.org/10.1007/s40745-019-00239-y
28. Prasetio, D., & Harlili, D. (2016). Predicting football match results with logistic
29. Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernández, J., & Medina,D. (2018). Effective injury forecasting in soccer with GPS training data and machine learning. PloS one, 13(7), e0201264. https://doi.org/10.1371/journal.pone.0201264
30. Şahin, M. . Optimization of dynamic ticket pricing parameters. Journal Of Revenue And Pricing Management, 18(4), 306-316. https://doi.org/10.1057/s41272-018-00183-1
31. Şahin, M. 8 . Dynamic pricing for sports events. Journal Of International Scientific Researches, 482-488. https://doi.org/10.21733/ibad.473973
32. Şahin, M. 8 . Forecasting attendance demand of sports games. Journal Of International Scientific Researches, 489-495. https://doi.org/10.21733/ibad.473975
33. Stein, M., Breitkreutz, T., Haussler, J., Seebacher, D., Niederberger, C., Schreck, T., ... & Janetzko, H. (2018, October). Revealing the invisible: Visual analytics and explanatory storytelling for advanced team sport analysis. In 2018 International Symposium on Big Data Visual and Immersive Analytics (BDVA) (pp. 1-9). IEEE.
34. Stensland, H. K., Gaddam, V. R., Tennøe, M., Helgedagsrud, E., Næss, M., Alstad, H. K., ... & Johansen, D. (2014). Bagadus: An integrated real-time system for soccer analytics. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 10(1s), 1-21.
35. Strobl, C., Malley, J. and Tutz, G. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification andregression trees, bagging, and random forests. Psychological Methods, 14, 4 (December 2009), 323–348. https://doi.org/10.1037/a0016973
36. Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley, Reading, MA. Exploratory data analysis. Addison-Wesley, Reading, MA.
37. Ulmer, B., Fernandez, M., & Peterson, M. (2013). Predicting soccer match results in the english premier league (Doctoral dissertation, Stanford University).
38. Yezus, A. (2014). Predicting outcome of soccer matches using machine learning. Saint-Petersburg University.
39. Zammit, M. (2018). Predictive Analysis of Football Matches using In-play Data, http://www.zammitmatthew.com/resources/msc-viva-dissertation.pdf.
40. Zawbaa, H. M., El-Bendary, N., Hassanien, A. E., & Kim, T. H. (2012). Event detection based approach for soccer video summarization using machine learning. International Journal of Multimedia and Ubiquitous Engineering, 7(2), 63-80.