RevistaEconomică
Journal of economic-financial theory and practice

Vol. 75, issue 2 Year 2023
DOI: 10.56043/reveco-2023-0013

ANALYSIS OF SPORTS PERFORMANCES USING MACHINE LEARNING AND STATISTICAL MODELS - A GENERAL ANALYSIS OF THE LITERATURE

Irina-Cristina COJOCARIU[1]

Alexandru Ioan Cuza University of Iasi, Romania, 0009-0008-1651-5670

Abstract:
The attendance of football fans at the matches played in the big city stadiums has a significant impact on the incomes of football clubs, an aspect studied more in recent years in the specialized literature, but with the summary presentation of related analysis techniques. Machine Learning remains certainly one of the preferred methodologies that has shown gratifying results in the fields of sports classification and prediction. One of the ever-growing fields that require good accuracy continues to be sports prediction, due to the huge amounts of money involved (player transactions, betting market, etc). These predictive models created for application within various clubs become a starting point for creating revenue maximization strategies. Taking into account these aspects, I will start by presenting the necessary steps in cleaning the data sets, continuing with the data preparation and their exploratory analysis by presenting the techniques offered by CRAN for the use of the R language and by non-programmers. Therefore, after the data set is prepared, we can start formulating the research questions, and this paper aims to present an objective analysis of the sports prediction models presented in specialized papers and the directions that can be followed in future research in the sports field, especially football.

Keywords: *Sports Performances, Statistical Models, Machine Learning*

JEL classification: *C53, L83, Z20*

1. Introduction

Numerous variables, including a team's past performance, past match results and individual player data, can be aggregated to determine the odds of winning or losing matches in the near future. The final decision regarding the probability of identifying the winning team is a very important one and requires the best possible precision due to the assets involved in the betting process; therefore, bookmakers, fans and potential sponsors are all interested in their favourite team's chances of winning. Therefore, the great challenge to obtain the most accurate results remains the concern over time for different actors (from researchers, investors, to the media).

The increasing demand for electronic (and often publicly available) data has led to a growing interest in the development of models and intelligent prediction systems for predicting match results. We have to keep in mind that most team and even player level data is lost because it is not stored electronically or made available to researchers to carry out research and analysis, of course for the benefit of the clubs.

Using structured experimentation approaches to the prediction problem for sports results will be helpful to get the best possible results from the data set. In the scientific paper of Bunker and Thabtah (2019) from the University of Japan, an intelligent architecture for predicting sports results is described, it proposes the steps for a machine learning framework and briefly describes the characteristic points of the data used to predict the results sports and how they fit into the presented framework.

Although the use of machine learning models for sports prediction is constantly increasing, there is still a need for more accurate models, and for this a more up-to-date, more comprehensive data set is required to obtain results that can have a significant impact. ML seems to be the most suitable method

[1] *cojocariu.irina96@yahoo.com*

for sports prediction, because using the right techniques it generates predictive models capable of predicting the results of matches based on predefined variables obtained from the analysis of previous results.

2. Methodology

The main purpose of the paper is to present two important learning methods in sports prediction (random forest and extreme gradient boosting), with an emphasis on the exploratory part of the data and related trends. The specialized literature presents predictive models that can be applied at a general level depending on the purpose of each one (anticipating the result, loss or win, predicting the number of supporters at the matches played in the stadium, etc.).

Using keywords relevant to sports prediction in the Google Scholar and Web of Science search engines, I was able to obtain the list of the most relevant papers on the subject at hand. Next, I will begin to introduce exploratory data analysis and cover two learning methods for prediction, Random Forest and XGBoosting.

3. Results and discussions

3.1. Literature review

In the vast field of sports analysis, this aspect of attendance has been analyzed for some time, especially in recent years. And, from the experts' criticism, a list of main variables can be obtained without which a quality model cannot be created. Regarding the aspect of the participation of fans in the matches played in the stadium, this is one of the main sources of obtaining direct income for sports teams, both at the level of large clubs and at the level of smaller clubs, and the theoretical and empirical research of public demand it is an integral part of the field of sports economics and sports management.

Since the 70s, studies have been carried out on the variables that influence in one way or another the presence at the stadium, Noll and Demmert, being among the researchers who undertook research to identify the determining factors. Each article, in addition to the basic variables, considered different types of variables that could influence the demand for tickets. The researchers also analyzed factors such as the income of the local population, the age of the stadium, the availability of replacements and the total population.

In an early study from 1975, the researchers Hart, Hutton and Sharton obtained a model for estimating the attendance of supporters at the stadium using the available data of four English Premier League clubs. An interesting variable was the geographical distance between the stadiums of the two teams. This proved to have a significant impact on the final results.

The researchers García and Rodríguez conducted a study in 2002 in which they analyzed the available data of the Spanish Football League to create an equation of attendance at the stadium. Among the variables of interest at that time we find the date on which the match took place, the weather conditions, and the quality of the participating teams (host and guest).

The research of the variables that influence the participation of supporters in the stadium continued, in this regard in 1992, Dobson and Goddard carried out a survey regarding the presence of fans at the matches played by the clubs of the English football league, and as variables of interest the ranking position of the host team was highlighted. Several analyzes were carried out based on the available data from Major League Baseball (Bruggink et al, 1996), (Rascher, 1999), (Forrest et al, 2002), all arguing that the models resulted that the variable showing the performance of the home team is more important than the performance of the away team.

Another variable with a significant impact on the participation of fans in the matches at the stadium was identified in the scientific articles as the number of points scored by the participating teams in the last 5 matches. And, in 2007, another article draws attention to the impact that important and well-known players have in increasing participation in matches (Brandes et al, 2007).

Şahin is one of the researchers who in recent years has analyzed this field of sports predictions, especially the dynamic analysis of ticket prices and implicitly the participation in the matches at the stadium. He considered in his studies since 2018 the analysis of the official websites of sports clubs from

Vol. 75, issue 2 Year 2023
DOI: 10.56043/reveco-2023-0013

different countries in order to create generally valid models. The goal being to have a significant impact on sports clubs regardless of their size or the country they belong to. His research thus becomes a valuable contribution in the promotion of sports analysis and prediction techniques.

3.2. Exploratory data analysis

Exploratory data analysis (EDA) emerged in the 1970s under the guidance of J.W. Tukey and refers to the initial investigation of data using statistical and visualization techniques (Tukey, 1977). EDA is a fundamental and primary step in any data analysis methodology. The competitive advantage of using the EDA task automation package is that users, even non-programmers, can achieve results with minimal code knowledge by applying the functions. In addition to these, we can mention the saving of time, the reuse of the code in different researches, but also the reduction of the rate of obtaining errors during the analysis.

In a study by Staniak and Biecek in 2019, it was found that among the EDA packages available on CRAN, DataExplorer (Cui, 2020) ranks second in user preferences (Staniak and Biecek, 2019). CRAN is a server that contains the R language code and documentation.

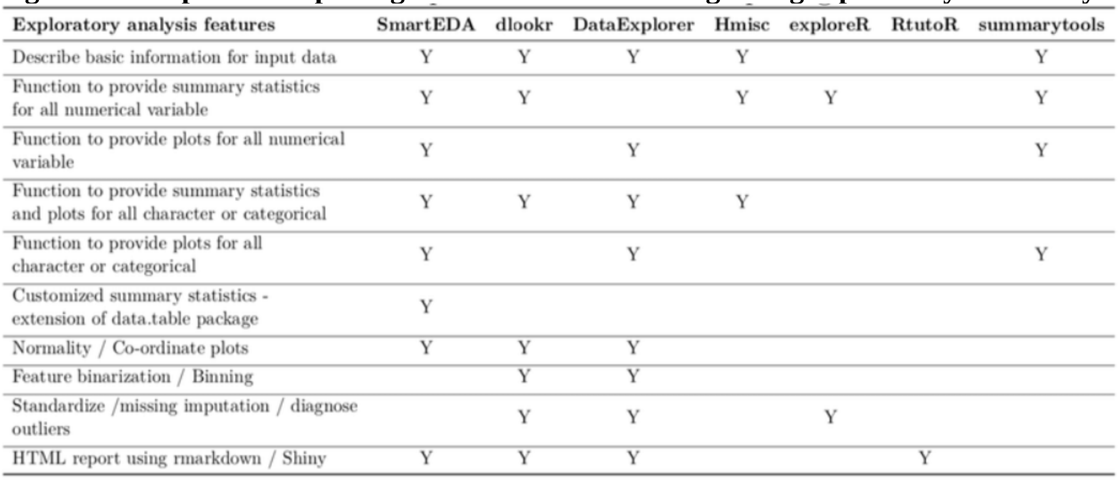
The main features of this package include creating a summary of the data set (size of the set, types of variables, missing values), identifying the type of missing values, plotting the distribution of variables and data transformations (replacing missing values, etc.).

Interest measurement for exploratory operations is a crucial component in many EDA systems, as well as for complete EDA automation (as described below). Many measures have been proposed for assessing the importance of analysis operations, each capturing a different facet of the broad concept (Greenwell et al., 2020). However, interest is often subjective (Chen et al., 2016) and changes dynamically, even within the same exploratory session (Patil, 2018). We reviewed two recent lines of research that use machine learning techniques to address this issue:

(1) Dynamic selection of measures of interest, where systems predict measures that most accurately capture the user, and

(2) Machine learning-based models for user interest, using, for example, active learning techniques (Demmert, 1973), (Hatcher, 2013) and rank learning (R Core Team, 2021).

Figure 1: Comparison of packages available in CRAN regarding exploratory data analysis



Exploratory analysis features	SmartEDA	dlookr	DataExplorer	Hmisc	exploreR	RtutoR	summarytools
Describe basic information for input data	Y	Y	Y	Y			Y
Function to provide summary statistics for all numerical variable	Y	Y		Y	Y		Y
Function to provide plots for all numerical variable	Y		Y				Y
Function to provide summary statistics and plots for all character or categorical	Y	Y	Y	Y			
Function to provide plots for all character or categorical	Y		Y				Y
Customized summary statistics - extension of data.table package	Y						
Normality / Co-ordinate plots	Y	Y	Y				
Feature binarization / Binning		Y	Y				
Standardize /missing imputation / diagnose outliers		Y	Y		Y		
HTML report using rmarkdown / Shiny	Y	Y	Y			Y	

Above is a comparison between the package called SmartEDA (Ubrangala et al., 2018) and other similar packages available on CRAN for exploratory data analysis, namely: dlookr (Ryu, 2018), DataExplorer (Cui, 2018), Hmisc (Harrell et al., 2018), exploreR (Coates, 2016), RtutoR (Nair, 2018) and resumetools (Comtois, 2018). The evaluation metric is the availability of various desired characteristics for performing exploratory data analysis, such as:

(a) Description of the basic information of the input data.

(b) Providing a feature.
(c) Summary statistics for all numerical variables.
(d) Providing graphs for all numerical variables.
(e) Providing summary statistics for all character or categorical variables.
(f) Providing graphs for all character or categorical variables.
(g) Custom summary statistics - data.table package extension.
(h) Normality / coordinate plots.
(i) Binarization / clustering of features.
(j) Standardization / absence of data completion / diagnosis of extreme values.
(k) HTML report using rmarkdown / Shiny.

We can see that the current version of SmartEDA has almost all the desired features mentioned above, except for points (h) and (i), i.e., the graphs and the binary normality function, respectively.

These two features will be incorporated in the next version and we are currently working on them. However, the most unique and powerful functionality offered by SmartEDA is point (f), i.e. the extension to the data.table package, which none of the other packages offer. Thus, SmartEDA adds value given the importance and popularity of data analysis among R users, especially for large datasets. Figure 1 shows that SmartEDA is better than almost all other packages available on CRAN. SmartEDA's closest competitor appears to be the DataExplorer package, but it lacks features (b) and (f), namely the function of providing summary statistics for all numeric variables, and the extension to the data.table package.

3.3. Models used in sports prediction

Random Forest (RF) and Extreme Gradient Boosting (XGB) are two highly popular machine learning techniques that build ensembles of classification or regression trees (Breiman et al., 1984). RF is an advancement of the sampling approach, constructing trees through randomized variable splits (Breiman, 2001), (Efron et al., 2016). XGB (Chen et al., 2016) is a more recent implementation of the gradient boosting framework (Friedman et al., 2000) and demonstrates excellent performance in both classification and regression tasks. RF typically reduces variance, while XGB excels at reducing bias (Efron et al., 2016).

The models are constructed following the tidy model framework (Kuhn et al., 2013), (Kuhn et al., 2020) utilizing the Ranger engine (Wright et al., 2017) for RF and the xgboost package (Chen et al., 2020) for XGB. The modeling process in Tidymodels leverages the doParallel package (Microsoft Corporation, 2020) to enhance speed. Variable importance is determined using the vip package (Greenwell et al., 2020). Since xgboost does not inherently support categorical predictors, these variables were transformed into dummy variables using the ::step_dummy option (Kuhn et al., 2020). In the presence of missing data, variable importance cannot be accurately estimated (Strobl et al., 2009); therefore, missing predictor values were imputed using the K-NN method available through the step_knnimpute option in tidymodels.

Both RF and XGB have hyperparameters that cannot be learned directly from data and require optimization (Luo, 2016), (Probst et al., 2019). In RF models, the tuned hyperparameters include the number of randomly selected attributes used for each node split and the minimum number of observations required in a node to proceed with the split. For the XGB models, the hyperparameters selected for tuning include the learning rate, the minimum loss reduction required to perform an additional split, the maximum tree depth, the size of random samples, and two additional parameters also modified in the random forest models.

In general, hyper-parameter tuning requires special expertise and many labor-intensive manual iterations (Luo, 2016), but this is currently not the case in ordered models, as different combinations of hyper-parameter values can be easily constructed (Kuhn et al., 2020). The risk of obtaining erroneous values can be further reduced by five-fold repeated cross-validation (Efron et al., 2020).

4. Conclusions

This paper introduces a framework for showcasing predictive models using the R language, wherein we have identified various packages that facilitate the data analysis process - encompassing data preparation, integration, processing, exploration, and declarative modeling (prediction) - at an advanced level.

An important aspect to consider is the increasing accessibility of Machine Learning to non-programmers as well. This accessibility ensures that the research findings can be consulted by both domain specialists and sports management personnel. We emphasize the application of the Pareto principle ("80/20 rule") in data analysis, wherein 80% of the time and effort is allocated to data set preparation, with only 20% being dedicated to its analysis.

The objective of this research is to present predictive models applicable in the domain of sports, while also being adaptable to other types of sports events.

The limitations we faced were primarily related to accessing the data and models utilized in sports analyses within the specialized literature.

For future research, the focus will be on developing predictive models and exploring the techniques employed at both the model and instance levels.

Acknowledgments

This paper is partially supported by the Competitiveness Operational Programme Romania under project number SMIS 124759 - RaaS-IS (Research as a Service Iasi).

References

- Bunker, R. P., & 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.
- Brandes, L., Franck, E. and Nüesch, S. (2007). Local Heroes and Superstars. *Journal Of Sports Economics*, 9(3), 266-286. https://doi.org/10.1177/1527002507302026.
- Bruggink, T. H., & Eaton, J. W. (1996). Rebuilding attendance in Major League Baseball: The demand for individual games. Baseball economics: Current research, 9-31.
- Chen T., Guestrin C.E. (2016). XGBoost: A Scalable Tree Boosting System. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*. ACM Press, 785–794. https://doi.org/10.1145/2939672.2939785.
- Cui B. (2020). DataExplorer: Automate Data Exploration and Treatment. *Rpackage version 0.8.2.* http://boxuancui.github.io/DataExplorer/.
- Demmert H.G. (1973). The economics of professional team sports. Lexington, Mass: Lexington Books.
- Dobson, S. M., & Goddard, J. A. (1992). The demand for standing and seated viewing accommodation in the English Football League. Applied Economics, 24(10), 1155-1163.
- Efron, B. (2020). Prediction, estimation, and attribution. International Statistical Review, 88, S28-S59.
- Forrest, D., Simmons, R. (2002). Outcome uncertainty and attendance demand in sport: the case of English soccer. *Journal Of The Royal Statistical Society: Series D (The Statistician)*, 51(2), 229-241. https://doi.org/10.1111/1467-9884.00314.
- García, J., & Rodríguez, P. (2002). The Determinants of Football Match Attendance Revisited: Empirical Evidence From the Spanish Football League. *Journal Of Sports Economics*, 3(1), 18-38. https://doi.org/10.1177/1527002502003001003.
- Greenwell B., Boehmke B. and Gray B. (2020). vip: Variable Importance Plots. R package version 0.2.2. https://CRAN.R-project.org/package=vip.
- Hatcher L. (2013). Advanced Statistics in Research. Shadow Finch Media, Saginaw, MI, USA

- Harrell, F. E., & Dupont, C. (2018). Hmisc: Harrell miscellaneous. R package version 4.1-1. R Found. Stat. Comput. https://CRAN. R-project. org/package= Hmisc (accessed 16 Feb. 2018).
- Hart R. A., Hutton J. and Sharot T. (1975). A statistical analysis of association football attendances. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(1), 17-27. https://doi.org/10.2307/2346700.
- Kuhn, M. (2020). FES: Code and Resources for Feature Engineering and Selection: A Practical Approach for Predictive Models by Kuhn and Johnson. GitHub Repository, available at https://github. com/topepo/FES, 308.
- Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., ... & Berk, M. (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. Journal of medical Internet research, 18(12), e323.
- Patil I. (2018). ggstatsplot: 'ggplot2' Based Plots with Statistical Details. CRAN. Retrieved from https://cran.r-project.org/web/packages/ggstatsplot/index.html.
- Prasetio, D., Harlili, D. (2016). Predicting football match results with logistic regression. *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*. https://doi.org/10.1109/icaicta.2016.7803111.
- R Core Team. (2021). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria*. https://www.r-project.org/.
- Staniak, M., Biecek, P. (2019). The landscape of R packages for automated exploratory data analysis. https://doi.org/10.48550/arXiv.1904.02101.
- Şahin, M. (2018). Dynamic Pricing For Sports Events. *Journal Of International Scientific Researches*, 482-488. https://doi.org/10.21733/ibad.473973.
- Şahin, M. (2018). Forecasting Attendance Demand Of Sports Games. *Journal Of International Scientific Researches*, 489-495, https://doi.org/10.21733/ibad.473975.
- Tukey, J. W., et al (1977). *Exploratory data analysis.*