Book title: BASKETBALL DATA SCIENCE

Authors:

Paola Zuccolotto – http://orcid.org/0000-0003-4399-7018

Marica Manisera – http://orcid.org/0000-0002-2982-0243

Chapter 1 – Introduction

Chapter 1 discusses some basic concepts about Data Science and Knowledge Representation, with specific reference to the context of sports analytics in general, and specifically in basketball. Data Science has not to be considered a substitute for human intelligence, but rather a tool that supports basketball technical experts in their choices and decisions by giving a quantitative answer to the questions they pose. After that, the most important scientific literature on basketball analytics is reviewed, with reference to the main topics covered, such as predicting the outcomes of games or tournaments, determining discriminating factors between successful and unsuccessful teams, examining the statistical properties and patterns of scoring during the games, analyzing a player's performance and the impact on his team's chances of winning, monitoring playing patterns with reference to roles, designing the kinetics of players' body movements with respect to shooting efficiency, depicting the players' pathways and trajectories, studying teams' tactics and identifying optimal game strategies, investigating the existence of possible referee bias, measuring psychological latent variables and their association to performance. Finally, the structure of the book is outlined.

Chapter 2 – Data and Basic Statistical Analyses

Chapter 2 gives a description of the most common types of basketball data, with specific reference to game statistics, box scores, play-by-play (event-log) and data detected by technological devices such as GPS sensors, wearable technologies or other player tracking systems. After that, a wide set of basic descriptive Statistics tools applied to basketball data is illustrated: Dean Oliver's four factors, pace, offensive rating and defensive rating computation, bar and radial plots built using game variables, percentages or other standardized statistics, scatter plots of two selected variables, bubble plots able to represent several features of players or teams in a unique graph, variability and inequality analysis (standard deviation, variation coefficient, range, Gini index, Lorenz diagram), shot charts with the court split into sectors colored according to a selected game variable and annotated with scoring percentages. All the topics are discussed by means of case studies and practical examples carried out using the open source R package `BasketballAnalyzeR`; codes are supplied in order to allow the reader to reproduce the presented analyses.

Chapter 3 – Discovering Patterns in Data

Chapter 3 discusses how Data Science can discover some kinds of patterns, i.e. intelligible regularities among variables, in order to reveal the hidden mechanisms governing the analyzed phenomena. Specifically, the Chapter deals with detecting statistical associations between variables (with a focus on linear correlation), building maps able to display individual cases according to their similarity by means of Multidimensional Scaling, analyzing network relationships (with specific reference to the assist-shot passing sequence), estimating the density of events (with respect to a covariate or in space, using density shot charts) and the joint density distribution of two variables. All the topics are discussed by means of case studies and practical examples carried out using the open source R package `BasketballAnalyzeR`; codes are supplied in order to allow the reader to reproduce the presented analyses. The final paragraph presents a structured case study concerned with identifying the factors determining high-pressure game situations and investigating how they affect the scoring probability, using Classification and Regression Trees (CART).

Chapter 4 – Finding Groups in Data

Chapter 4 deals with finding some hidden framework according to which data could be reorganized and categorized by means of Cluster Analysis algorithms, methods and rules aimed to assign instances into classes, which are not defined a priori, and which are supposed to somehow reflect the structure of the entities that the data represents. After a brief recall of the main Cluster Analysis algorithms (namely, k-means and hierarchical agglomeration), applications in basketball are examined thanks to three examples dealing with clustering of teams, shots and players, respectively. Each example also discusses the issues of deciding the optimal number of clusters, measuring the quality of the clusterization by means of the explained variance, observing the cluster tree (for the hierarchical algorithm) and interpreting the cluster profiles. Further analyses within the obtained clusters are also proposed with bar plots, bubble plots, Lorenz and variability diagrams. All the topics are discussed by means of case studies and practical examples carried out using the open source R package `BasketballAnalyzeR`; codes are supplied in order to allow the reader to reproduce the presented analyses. The final paragraph presents a structured case study concerned with new basketball roles definition according to the players' game statistics, using an unsupervised Artificial Neural Network called Self-Organizing Map (SOM) and some fuzzy clustering procedures.

Chapter 5 – Modeling Relationships in Data

Chapter 5 addresses statistical models, tools and procedures aimed to approximate the mechanisms or the rules that govern the functioning of phenomena. After a brief introduction about the two approaches to statistical modelling, i.e. the data modelling and the algorithmic modelling culture, some specific methods are described, namely a linear model (simple linear regression) and two nonparametric regression techniques (polynomial local regression and Gaussian kernel smoothing). Two specific insights focus on estimating the scoring probability and the expected number of points scored as a function of some game variables such as the period time (seconds played in a given quarter), the total time (seconds played in total), the play length (time between the shot and the immediately preceding event), the shot distance (distance in feet from the basket). All the topics are discussed by means of case studies and practical examples carried out using the open source R package `BasketballAnalyzeR`; codes are supplied in order to allow the reader to reproduce the presented analyses. Finally, a paragraph shows a structured case study dealing with the analysis of surface areas dynamics and their effects on the team performance, using data recorded with player tracking systems. The dynamics are analyzed with Markov Switching Model and Vector Auto Regressive models.

Chapter 6 - The R package BasketballAnalyzeR
Marco Sandri – https://orcid.org/0000-0002-1422-5695

Chapter 6 describes the basic features of the `BasketballAnalyzeR`, an open-source package for the statistical computational language R, designed for the analysis and visualization of basketball data and specifically addressed to the readers of this book. `BasketballAnalyzeR` takes advantage of the powerful graphical abilities added to R by the `ggplot2` package and allows to produce publication-quality graphics with minimal effort, with the two main guiding principles being simplicity and flexibility.

The webpage `bdsports.unibs.it/basketballanalyzer` is devoted to the package and constantly updated with news and upgrades. The Chapter deals with preparing data (how to manipulate data frames to be used with `BasketballAnalyzeR`), customizing plots (how to annotate graphs, add titles, change axes labels, range and ticks, set panel background and border colour, arrange into grids, …) and building interactive graphics by means of the powerful R graphing library `plotly`. All the topics are discussed by means of practical examples; codes are supplied in order to allow the reader to reproduce the presented analyses.