Original research

International Journal of Sports Science & Coaching

International Journal of Sports Science & Coaching
2018, Vol. 13(4) 569–589

DOI: 10.1177/1747954117737492
journals.sagepub.com/home/spo
SAGE

# Big data analytics for modeling scoring probability in basketball: The effect of shooting under high-pressure conditions

**Paola Zuccolotto[1], Marica Manisera[1] and Marco Sandri[2]**

**Abstract**

In this paper, we analyze the shooting performance of basketball players by examining the factors that may generate high-pressure game situations. Using play-by-play data from the Italian "Serie A2" Championship 2015/2016 to build the model, we validate the main results using data from the Olympic Basketball Tournament "Rio 2016" to determine whether the relationships we identified can be confirmed using data from players at a very different professional level. After a preliminary exploratory analysis, we (1) develop a multivariate model based on the Classification and Regression Tree algorithm in order to investigate how selected high-pressure situations, jointly considered, affect scoring probability and then propose new shooting performance measures; (2) investigate players' personal reactions to selected high-pressure game situations by introducing additional new measures, improving the indices currently used to measure shooting performance. The results are interesting and easy to interpret with the aid of some insightful graphical representations. Our approach can be exploited by both scouts and coaches to understand important player characteristics and, ultimately, to measure and enhance a team's performance.

**Keywords**

Basketball analytics, performance analysis, statistical model

## Introduction

In the past few decades, the application of statistical thinking to sports has rapidly gained interest, as documented by the wide variety of scientific research on this theme as well as the publication of some insightful collections of statistical analyses applied to data from a broad range of sports.[1–3]

For studies of basketball, several statistical techniques have been applied with a wide variety of different aims, ranging from simply depicting the main features of a game by means of descriptive statistics[4] to the investigation of more complex problems, such as forecasting the outcome of a game or a tournament,[5–12] analyzing players' performance,[13–19] studying the network of players' pathways from the in-bounds pass to the basket[20] and their spatial positioning,[21] or identifying optimal game strategies.[22]

The range of possible questions that may be answered by statistical analysis is growing due to the online availability of large sets of play-by-play data and increasing computational power, allowing big data analysis.[23,24] From a methodological point of view, a huge amount of raw data available on actions and strategies, combined with the absence of a sound theory explaining the relationships between the variables involved, make these questions a challenge for data scientists.

The aim of this paper is to analyze the shooting performance of basketball players by taking into account the factors that may generate high-pressure game situations.

In doing so, we must be aware of some issues.

One important consideration is the definition of a "high-pressure" game situation. As argued by Tango et al.,[25] who examined clutch players in baseball, there is no concrete definition for a high-pressure situation. As we will explain in detail below, several

Reviewers: Andrea Bonanomi (Catholic University of the Sacred Heart, Italy)
Albrecht Zimmerman (University of Caen, France)

[1]Department of Economics and Management, University of Brescia, Brescia, Italy
[2]DMS StatLab, University of Brescia, Brescia, Italy

**Corresponding author:**
Marica Manisera, Department of Economics and Management, University of Brescia, C.da S. Chiara, 50, 25125, Brescia, Italy.
Email: marica.manisera@unibs.it

attempts have been made to resolve this problem, but researchers are still far from finding an acceptable solution. An interesting analysis by Jones[26] recognizes that a high-pressure competitive situation is not just a psychological issue, as it may derive both from competition-related factors (external) and anxiety (internal). Expanding on this idea, in this work, we identify a game situation as being ''high-pressure'' if it is for some reason more troublesome and demanding than at other points in the game. With this definition, we do not make any distinction between external and internal pressure factors, which is clearly a drawback of the analysis; however, it would be realistically impossible to make that distinction, as often highlighted in the literature on this topic. In fact, Apesteguia and Palacios-Huerta[27] point out that in competitive environments, phenomena are typically too complex to be empirically tractable in a way that allows one to discern psychological elements from within the complex behavior exhibited by humans, as the difficulties in clearly observing actions, outcomes, choices of risky strategies, and other relevant variables are exceedingly high. Goldman and Rao,[28] who work on this theme in the context of basketball, observe that play-by-play data allow observational studies and do not permit investigation of the complex set of all possible hypotheses. Therefore, we will limit ourselves to observations of what happens in selected game situations that are more troublesome and demanding, without venturing any conjecture or causal inference on whether the performance is affected by external factors, psychological factors, or both (the last being, however, the most likely option).

In terms of statistics, three main challenges must be taken into account. First, a large amount of play-by-play data are needed in order to obtain robust estimates of the scoring probabilities. Second, the analyzed phenomenon is characterized by complex relationships and interactions among the variables, so we propose the use of multivariate data mining tools that are able to deal both with data complexity and large datasets. Third, in analyzing high-pressure game situations, we have to consider that, according to some psychological studies,[29,30] some athletes view competitive situations as challenging, and others perceive the same situations as stressful and anxiety-provoking. This is another important point to stress: we do not assume a priori that a high-pressure game situation will necessarily result in a lower success probability, in agreement with the definition of ''clutch skill'' as a player's ability to perform better in ''clutch'', or high-pressure situations.[25] For this reason, it may be difficult to detect by statistical analysis an effect of high-pressure situations from large datasets including several players, as the overall average performance may remain unchanged if some players improve their performance while others get worse. To cope with this possibility, we will carry out the analysis considering both the average from multiple players and the personal characteristics of individual players.

Our analysis will be carried out through two consecutive steps, with the following intermediate aims:

Step 1. To develop a model describing the impact of some selected high-pressure game situations on the probability of scoring. This model should allow us (a) to determine which situations have, on average, an actual effect on the probability of scoring and, on the other hand, which situations do not have any average effect, typically because they are managed differently by different players and (b) to provide new shooting performance measures, improving upon currently employed shooting statistics.

Step 2. To assess players' personal reactions to selected high-pressure game situations judged by basketball experts as particularly important.

The paper is organized as follows. In Section Data, the main characteristics of the analyzed data are presented. In Section ''Modeling the impact of high-pressure game situations on the shooting performance'', after showing some preliminary univariate analyses, we present the multivariate model – built using a popular data mining tool – describing the impact of selected high-pressure situations on scoring probability and the new performance measures based on this model. Subsection ''Comparison to Rio16 Data'' is devoted to comparing these results with those obtained performing the same analyses on a different dataset. Section ''Players' personal reactions to selected high-pressure game situations'' presents the analysis of players' personal reactions to high-pressure situations. In the Concluding remarks, we present our conclusions from the data analysis.

## Section "Data"

The analysis is performed on a dataset derived from the play-by-play records of all matches played during the Italian ''Serie A2'' Championship 2015/2016. ''Serie A2'' is the second tier of the Italian league pyramid, just below the first division, ''Serie A'', and it is the highest Italian non-professional competition level. This dataset, hereafter referred to as A2Ita, is obtained by extracting from play-by-play data all of the attempted shots, described according to all of the information that can be extrapolated from the data (e.g. the player, the time, the shot clock, the score, and other variables that will be described later). Data were retrieved from www.legapallacanestro.com (A2Ita), where all of the matches' actions are recorded with the approval of the Italian Basketball Federation (FIP). This dataset contains almost 70,000

**Table 1.** Main features of datasets A2ITA and RIO16.

| Dataset | A2ITA | RIO16 |
|---|---|---|
| Competition | Championship – regular season | Olympic tournament |
| Period | 4 Oct 2015–23 April 2016 | 6–21 Aug 2016 |
| Gender | Male | Male |
| Number of matches | 480 | 38 |
| Number of teams | 32 | 12 |
| Number of players | 438 | 144 |
| Number of 2-point shots | 33,682 (48.3%, 50.9% Made) | 3,101 (47.9%, 52.2% Made) |
| Number of 3-point shots | 21,163 (30.4%, 34.1% Made) | 1,780 (27.5%, 33.8% Made) |
| Number of free throws | 14,843 (21.3%, 73.5% Made) | 1,589 (24.6%, 74.8% Made) |

shots, and hence the sample size is large enough to guarantee robust estimates of the scoring probabilities, even in situations that may occur only occasionally. It is worth considering that the reaction to pressure may be different according to the professional level of the players, as noted by Madden et al.[31] To take this element into account, the most important results of Step 1 will then be checked on the smaller dataset from the Olympic Basketball Tournament "Rio 2016," referred to as RIO16, which was retrieved from the official website of the International Basketball Federation (FIBA), www.fiba.it. Table 1 summarizes the main features of the two datasets.

## Modeling the impact of high-pressure game situations on the shooting performance (Step 1)

In this section, we estimate how the scoring probability is affected by different game situations. We first carry out univariate preliminary analyses using a traditional statistical method, such as non-parametric regression via kernel smoothing. Then, we present a multivariate analysis using a popular data mining tool. Finally, we introduce new player performance measures based on the results we obtained. In the end, we validate the most important results on the RIO16 dataset, collected from a tournament with players of a profoundly different professional level.

As pointed out in the Introduction section, the first problem we face here is the definition of which situations we consider to be of "high-pressure." From a game-related perspective, Tango et al.,[25] opted for a general definition invoking the idea of the game being on the line ("*We'll define a high-pressure situation as one in which runs are needed in the very near future but the game is not yet out of hand*"), while Goldman and Rao[28] who analyzed National Basketball Association (NBA) data quantified pressure as the marginal impact of an additional point to be modeled as a function of score margin and time remaining with the help of a Probit model. On the other hand, focusing on psychological issues[30] identified 20 stressful situations (using their Stressful Situations in Basketball Questionnaire, SSBQ), entirely composed of game states relating to a range of offensive, defensive, and neutral situations occurring in competitive basketball. The six categories of stressful situations were classified by the authors as: Being Outplayed, Errors in Personal Skills, Errors in General Play, Game Tension, Team Performance, and Other Performance.

Here, we identify a game situation as being "high-pressure" if it is for some reason more troublesome and demanding, without trying to distinguish whether the pressure comes from game-related factors, psychological factors, or both. Following this general definition and considering the available data, we identified, with the help of basketball experts, four main types of situations that may generate pressure on the player when a shot is attempted:

(a) when the shot clock is going to expire;
(b) when the score difference with respect to the opponent is small;
(c) when the team as a whole has performed poorly during the match up to that particular moment in the game;
(d) when the player has missed his previous shot.

It is worth noting that, consistent with our previous claim, all of these situations can be considered "high-pressure," due to both game-related and psychological factors. Just to mention a couple of examples, when the shot clock is going to expire, the player is usually subject to both a higher defensive pressure from the opponent and to anxiety due to the necessity to hurry up; when the team is performing poorly, pressure may derive both from the presence of a strong opponent and the sense of urgency.

The time to the end of match (time left on the clock) may also be relevant, but it is reasonable to assume

that it only generates pressure in interaction with other variables, for example, the score difference (as demonstrated by Goldman and Rao[28]). For this reason, we do not take into account this variable in the univariate preliminary analyses, but we will include it in the multivariate study presented in the next section.

The variables for the four high-pressure game situations we identified were:

(a) SHOT.CLOCK: numerical variable, ranging from 0 to 24 s, denoting the time on the shot clock at each attempted shot;
(b) SC.DIFF: numerical variable denoting the score difference with respect to the opponent when each shot is attempted;
(c) MISS.T: numerical variable, ranging from 0 to 1, denoting for each match the fraction of missed shots for the whole team up to the moment when each shot is attempted;
(d) MISS.PL: categorical variable denoting whether the previous shot by that player scored a basket ("Made") or not ("Missed").

## Preliminary analyses

For Cases a, b, and c, we performed univariate nonparametric regressions via kernel smoothing on the dependent variable MADE, assuming values 1 and 0 according, respectively, to whether the attempted shot scored a basket or not. Variables SHOT.CLOCK, SC.DIFF, and MISS.T acted as covariates. In order to investigate the variability of the estimated relationships, we used data from 1000 bootstrap samples of size $n_{boot} = 5000$ for each type of shot (2-point or 3-point).

Figure 1 displays the results obtained for Case a (covariate SHOT.CLOCK), separately for 2-point and 3-point shots (the shot clock cannot be considered a source of pressure for free throws). The colored areas range from the 5th to the 95th percentile of the bootstrap estimates of the scoring probability in correspondence to each value of SHOT.CLOCK. The horizontal red-dashed lines denote the 2-point (left panel) and 3-point (right panel) field goal percentages, and the gray lines represent the estimated density function of shots for each value of SHOT.CLOCK, whose measure is reported on the right-hand side vertical axis. The shot clock exhibits a significant effect on the scoring probability, which appreciably decreases as the buzzer sound is closer. In the case of 2-point shots, shots attempted in the first seconds of the possession have the highest probability of scoring, as they usually occur in the case of a fast break. Conversely, for 3-point shots, early shots are less successful, as they usually occur when trying an end-of-quarter buzzer beater (often "from downtown", in basketball jargon). In any case, the density functions (gray lines) show that most of the attempted shots occur in the last 12 s.

It is worth pointing out that the variable SHOT.CLOCK describes the seconds until the buzzer sounds (i.e. the value displayed on the shot clock, regardless of whether the shot clock had been previously reset to 14 s). As a matter of fact, the extra time with respect to the 24 s may temporarily modify the pressure of the game and affect the scoring probability.

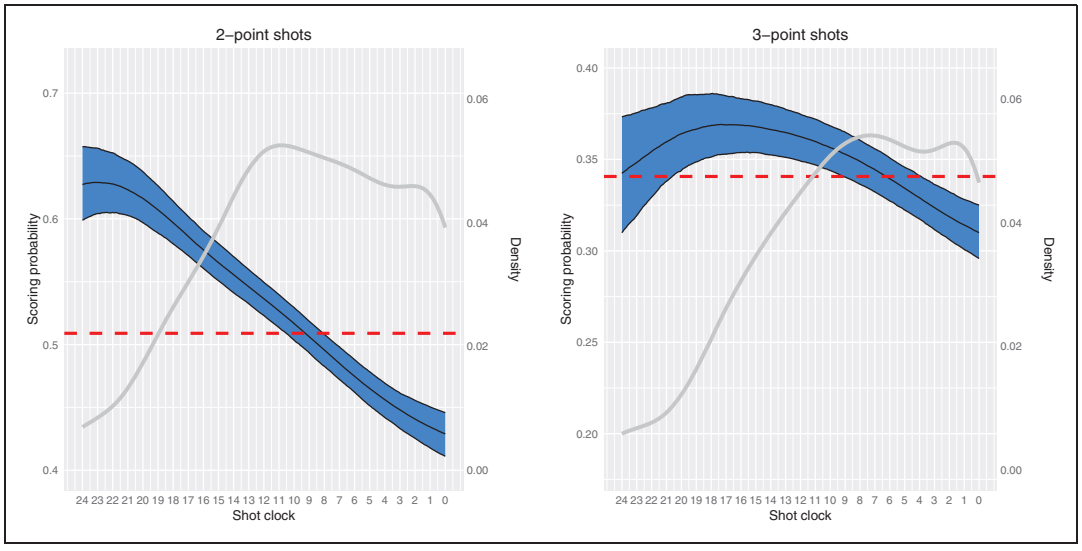


**Figure 1.** Bootstrap kernel estimates of the scoring probability conditional to SHOT.CLOCK (coloured areas) and distribution of shots by SHOT.CLOCK (gray lines), dataset A2ITA.

Figure 2 shows a separate kernel estimation of the scoring probability for game play after resetting the shot clock to 14 s (green lines superimposed on the previous graphs). The scoring probability still decreases as the shot clock runs down; nevertheless, it is considerably higher for 2-point shots and 3-point shots attempted in the very last seconds. In the multivariate analysis in Step 2, we will analyze more deeply the effect of resetting the shot clock to 14 s.

Similar to Figure 1, Figures 3 and 4 display the results for Cases b and c for 2-point shots, 3-point shots, and free throws.

In these two cases, the data do not exhibit any noteworthy direct relationship between the covariates and the scoring probability. Even in the absence of a univariate association between the variable MADE and the two covariates SC.DIFF and MISS.T, we cannot exclude the possible existence of a multivariate relationship, for example, in interaction with other variables. For this reason, we will include the two covariates in the analysis presented in the next section. Another important point is that an association could be detected at the player level. Figures 5 to 7 show the kernel estimation of the scoring probabilities for a selected player (orange lines superimposed on the previous graphs), Terrence Terrell Roderick Jr., who played for the Team "BCC Agropoli" and attempted 307 2-point shots, 152 3-point shots, and 125 free throws in the analyzed season.

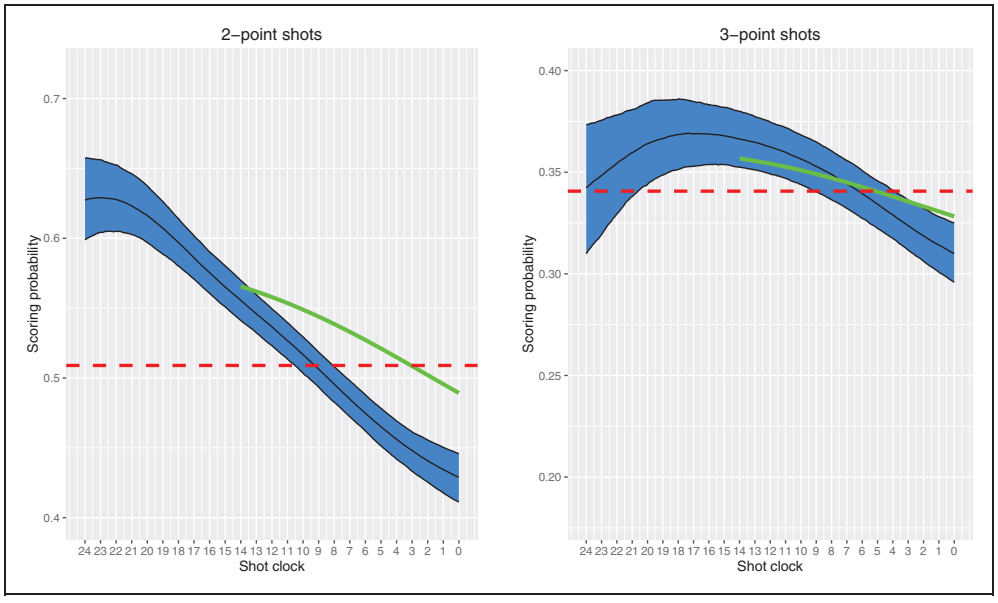


**Figure 2.** Kernel estimates of the scoring probability conditional to SHOT.CLOCK, with separate analysis of 14-s possessions (green lines), dataset A2ITA.

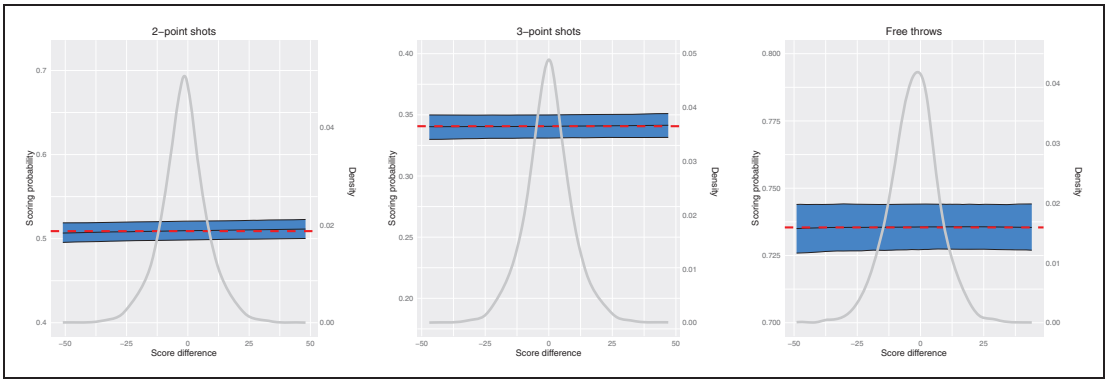


**Figure 3.** Bootstrap kernel estimates of the scoring probability conditional to SC.DIFF (coloured areas) and distribution of shots by SC.DIFF (gray lines), dataset A2ITA.

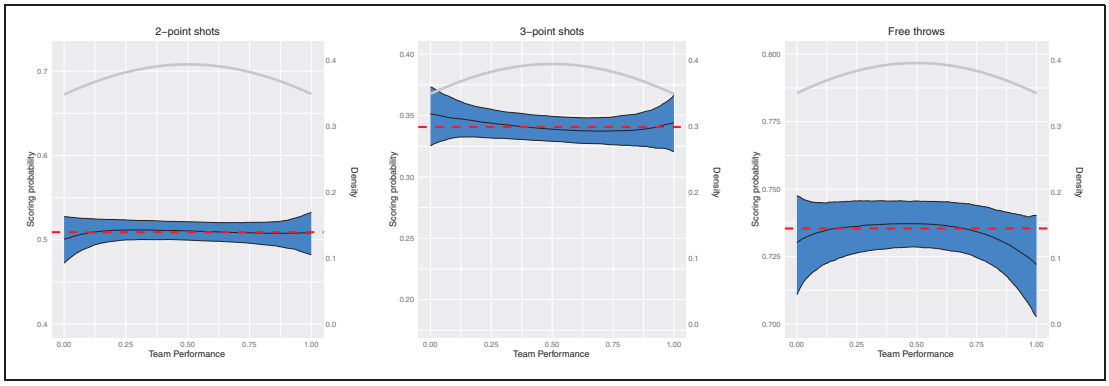


**Figure 4.** Bootstrap kernel estimates of the scoring probability conditional to MISS.T (coloured areas) and distribution of shots by MISS.T (gray lines), dataset A2ITA.

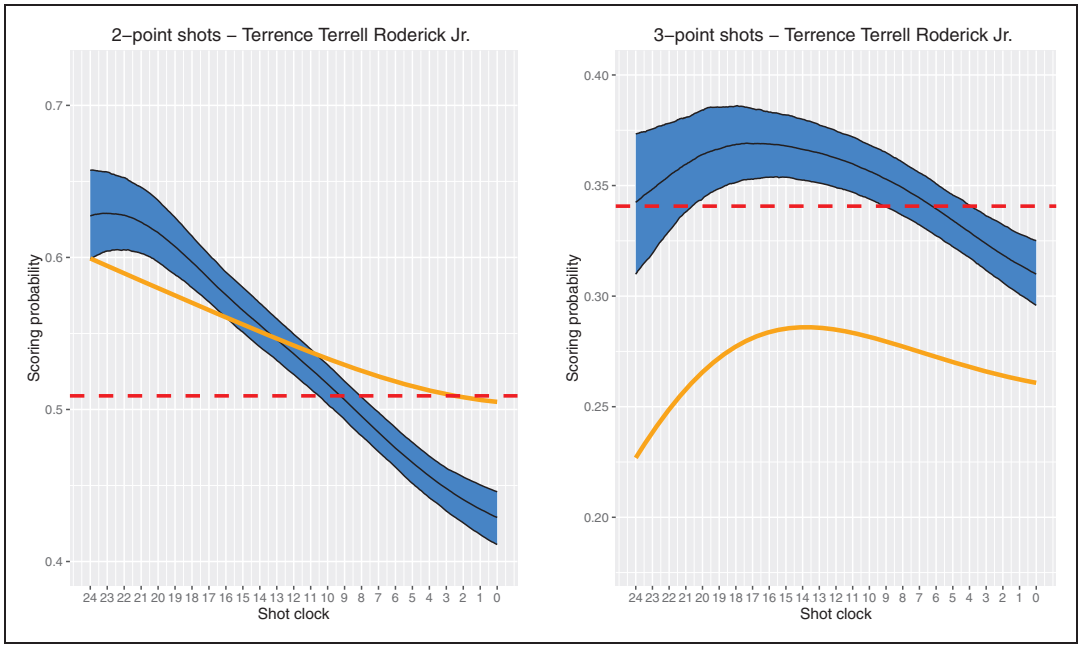


**Figure 5.** Kernel estimates of the scoring probability conditional to SHOT.CLOCK, with separate analysis of the player Terrence Terrell Roderick Jr. (orange lines), dataset A2ITA.

Interesting conclusions could be drawn comparing the performance of the player to the overall results: Terrence Terrell Roderick Jr. tends to perform better than the average both in terms of 2-point shots (especially when approaching the shot clock buzzer sound) and free throws, but his scoring probability decreases when the score margin is negative; he performs worse than the average for 3-point shots (with the same pattern as the overall results with respect to the shot clock), and he does even worse when the score margin is positive. Again, no association is detected with the variable MISS.T.

As concerns Case d, we estimated the scoring probability conditional to the two categories of MISS.PL (“Made” and “Missed”) on each bootstrap sample, and the boxplots of these estimates are shown in Figure 8.

Covariate MISS.PL is definitely associated with scoring probability in the case of free throws, with an appreciably higher scoring probability when the previous shot scored a basket. Nevertheless, the opposite relationship emerges for 3-point shots and, more weakly (with almost overlapping boxplots), for 2-point shots, where the probability of a successful shot

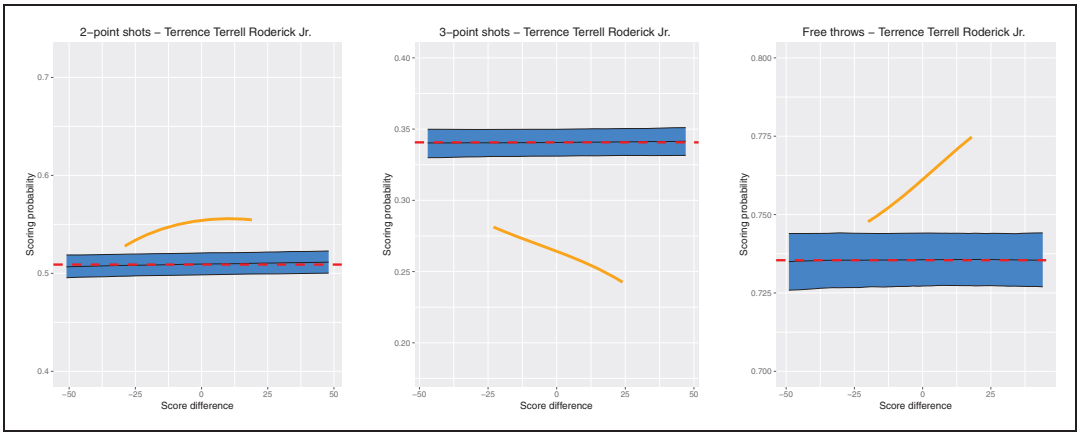


**Figure 6.** Kernel estimates of the scoring probability conditional to Sc.Diff, with separate analysis of the player Terrence Terrell Roderick Jr. (orange lines), dataset A2Ita.

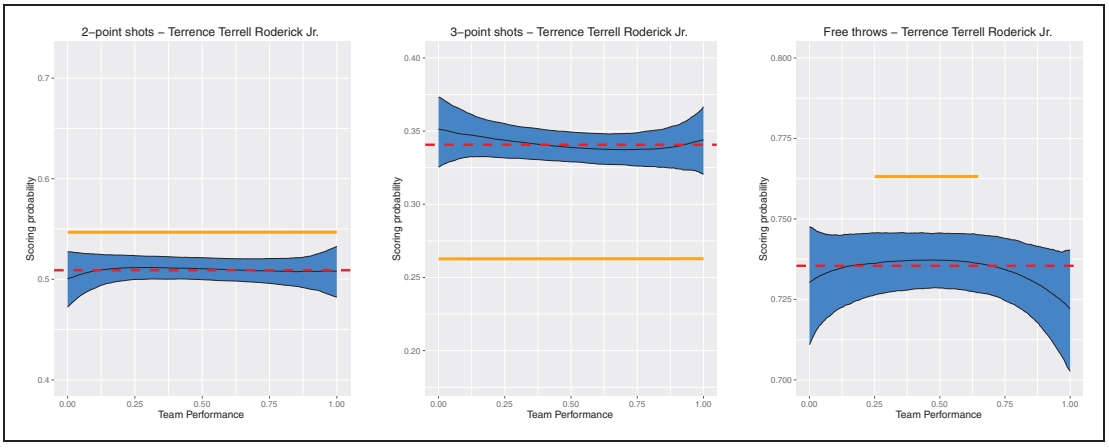


**Figure 7.** Kernel estimates of the scoring probability conditional to Miss.T, with separate analysis of the player Terrence Terrell Roderick Jr. (orange lines), dataset A2Ita.

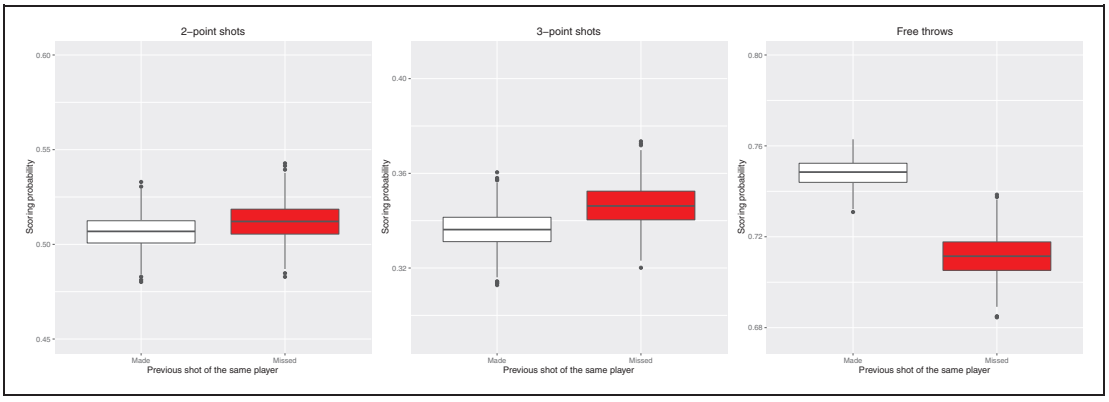


**Figure 8.** Boxplots of the bootstrap estimates of the scoring probability conditional to Miss.Pl (white: previous shot made, red: previous shot missed), dataset A2Ita.

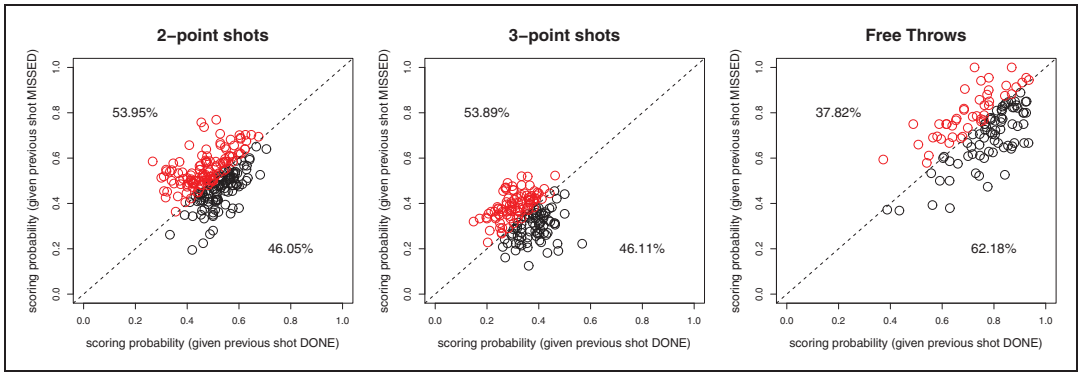


**Figure 9.** Players' scoring probabilities conditional to MISS.PL, dataset A2ITA.

appears to be higher when the previous shot was missed. This apparently odd result may be explained by recalling the two-fold nature of high-pressure situations, which are composed of both game-related and psychological factors. With free throws, we can confidently assume a strong impact of psychological factors,[32] and a success with the previous shot may improve the player's confidence. On the other hand, game-related factors may be more determinant for 2- and 3-point shots. For example, it is possible that, if a player makes a basket, he will be defended more aggressively, making the next shot harder, and vice versa if the player misses. For this case, we perform a deeper investigation at the player level in order to take into account the possible effects of the individual's performance (a bad shooter, for example, is more likely to repeatedly miss).

In Figure 9, the players are represented as points in a two-dimensional space, with $x$ and $y$ axes representing the scoring probability if the previous shot was made or missed, respectively. For each type of shot, the graph only accounts for players who have attempted at least 50 shots. The black points lying below the dashed line are players who tend to perform better when the previous shot was made. The graph confirms at the player level what we already observed at the aggregate level: for free throws, the majority of the analyzed players (62.2%) tend to perform better after a good shot. The opposite, but with smaller differences in the percentages, is the case for 2- and 3-point shots.

## *Multivariate analysis with Classification and Regression Trees*

The multivariate analysis allows us to assess the impact of the high-pressure situations we identified on the scoring probability by taking into account all of the joint associations among the variables. Due to the complex relationships characterizing the analyzed phenomena, the use of data mining techniques appears to be particularly appropriate (examples of data mining in sports can be found in literature[33,34]). We used Classification and Regression Trees (CART[35]) in order to take into account possible interactions among variables, as the preliminary analyses showed that direct relationships between the single covariates and the scoring probability were in some cases weak or absent. In addition, CART, which has recently be judged to be one of the top 10 algorithms in data mining,[36] allows us to obtain results characterized both by accuracy and easy interpretability, provided the sample size is large enough to avoid the well-known instability problems.[37] Given the large size of our dataset, interpretability is the main reason we prefer a single tree with respect to machine learning tools such as Random Forest,[38] Gradient Boosting Machine,[39] or other ensemble learning algorithms. In fact, these tools are strongly recommended when the sample size is small, because of their stability and accuracy, but, on the other hand, they are "black boxes."

We introduce into the analysis four additional variables: SHOT.TYPE (a categorical variable describing the type of shot, with categories "2P," "3P," and "FT" representing 2-point shots, 3-point shots, and free throws), TIME (a numerical variable accounting for the time to the end of each quarter), POSS.TYPE (a categorical variable describing whether the shot is made during the original 24 s on the shot clock or after the shot clock has been reset to 14 s), and QUARTER (a categorical variable indicating the game quarter). Finally, we model the dependent variable MADE using as covariates the numerical variables SHOT.CLOCK, SC.DIFF, MISS.T, TIME, and the categorical variables MISS.PL, SHOT.TYPE, POSS.TYPE, and QUARTER. One way to control the trees' instability and improve interpretability is to use categorical covariates instead of numerical ones, as this prevents trees from growing too deeply and focusing only on the most interesting situations.

To convert the numerical covariates into categorical ones, we consider two different approaches:

- Relying on experts' suggestions as to which situations should be considered "high-pressure," which include focusing on (1) the last 1–2 s of possession, very close to the shot clock buzzer sounding, (2) games where the score difference is low, for example, between −4 and 4, and (3) the last 1–2 min of each quarter (especially the final quarter).
- Using machine learning tools such as CART to determine optimal thresholds.

Of course, both approaches have pros and cons. The disadvantages of a pure machine learning approach include:

- A large sample size is able to prevent the predictions' instability, but small changes in the data may still result in different thresholds.
- The algorithm determines the split thresholds with the criterion of maximizing the heterogeneity reduction in the outcome variable. This leads to an automatic identification of high-pressure game situations, sometimes difficult to see by experts. The resulting splits comply with the algorithm objective but, in practice, might not help the aim of identifying high-pressure situations, when, for example, the resulting combination of covariates and thresholds cannot be used as a practical definition of high-pressure conditions.

For these reasons, we suggest identifying the cut-off values to convert numerical covariates into categorical ones by mixing the two approaches, combining the results of a machine learning procedure and the experts' suggestions. Nevertheless, we grew the tree using all of the numerical covariates (without transforming them into categorical ones), and the two approaches lead to very similar results: the resulting splits substantially overlap those obtained by the procedure proposed below, mixing machine learning with experts' suggestions.

As for the machine learning procedure, we propose adapting the idea of a variable importance measure to obtain a *threshold importance measure*. We recall that the introduction of a variable importance measurement in tree-based predictors dates back to Breiman et al.,[35] where the importance of a covariate in a tree-based prediction algorithm is defined as the total decrease of heterogeneity of the response variable when the feature space is partitioned recursively. The variable importance measure is then obtained by summing up all of the decreases of the heterogeneity index allowed by a given covariate in all of the nodes of the tree. Here, we propose to do the same at the threshold level. In detail, we proceed through the following steps:

1. We grow a tree to model the dependent variable MADE, using as covariates the numerical variables SHOT.CLOCK, SC.DIFF, MISS.T, and TIME, and the categorical variables MISS.PL, SHOT.TYPE, POSS.TYPE, and QUARTER. The tree is grown with a stopping rule based on a minimum leaf size (500) in order to limit the number of uninformative splits[40,41] without pruning.
2. For each numerical covariate, we consider all of the thresholds used in the tree to split nodes.
3. We sum up all of the decreases in the heterogeneity index allowed by each threshold of each covariate.

The results are displayed in Figure 10.
In detail, we conclude as follows:

- The variable SHOT.CLOCK presents three clear thresholds at 2, 10, and 17 s, which is partly consistent with the experts' suggestions of isolating the extremes.
- For the variable TIME, the thresholds' importance begins to grow in the last 100 s, which is also consistent with the experts' opinions regarding the relevance of the last 1–2 min of each quarter.
- The variable MISS.T presents thresholds between 0.34 and 0.58, but we do not recognize peaks that may suggest a possible best choice within this interval, so we may simply split according to quantiles, provided they lie in the interval [0.34,0.58].
- The variable SC.DIFF presents several different peaks. Among them, we recognize the possibility of following the experts' suggestions to isolate low score differences (between −5 and 1), but we also find some possible thresholds at the values −15 and 6.

On the whole, we decided to convert numerical variables into categorical variables according to the criteria summarized in Table 2.

The CART model with all of the categorical covariates was grown using the Gini index as the split selection criterion. After pruning, we obtained the tree as shown in Figure 11.

The goodness-of-fit, assessed with the AUC (area under the ROC curve) index, is 0.671 (with 95% confidence interval [0.667;0.675])[42]. Although the AUC value is only moderately satisfactory, all of the probability estimates are considerably robust, as the leaf-node size ranges from 2,681 to 13,031, with the exception of one leaf node (the one with an estimated scoring probability of 0.4678) having a size of 652. The tree provides very interesting interpretations about the impact of high-pressure game situations on the

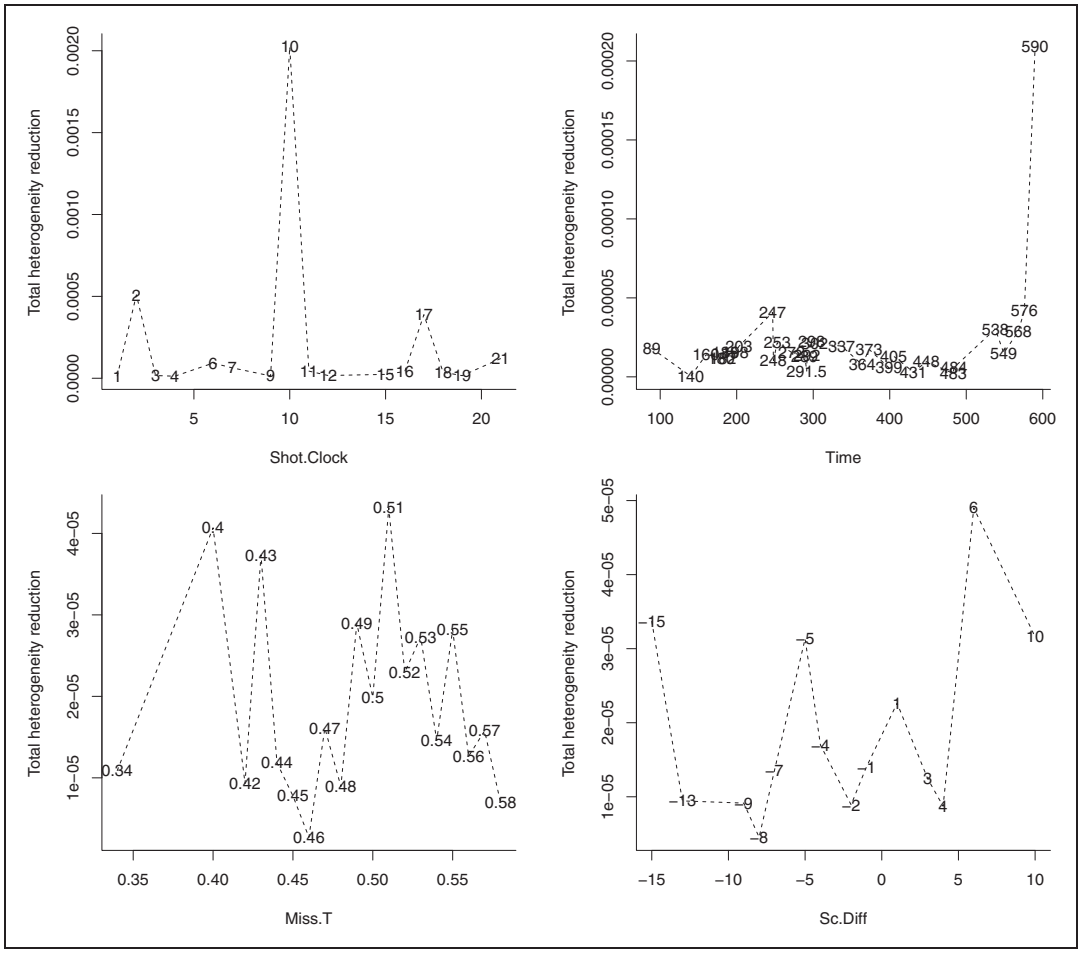


**Figure 10.** Thresholds importance measures, dataset A2ITA.

**Table 2.** Conversion into categorical covariates, dataset A2ITA.

| | |
|---|---|
| SHOT.CLOCK | early: SHOT.CLOCK> 17 |
| | early-middle: 10 < SHOT.CLOCK≤ 17 |
| | middle-end: 2 < SHOT.CLOCK≤ 10 |
| | time-end: SHOT.CLOCK≤ 2 |
| TIME | normal: TIME≤ 500 |
| | quarter-end: TIME> 500 |
| MISS.T | Bad: MISS.T ≤ 0:44 (25th percentile) |
| | Medium: 0:44 < MISS.T ≤ 0:56 |
| | Good: MISS.T > 0:56 (75th percentile) |
| SC.DIFF | less than -15: SC.DIFF ≤ −15 |
| | between −15 and −5: −15 < SC.DIFF ≤ −5 |
| | between -5 and 1: −5 < SC.DIFF ≤ 1 |
| | between 1 and 6: 1 < SC.DIFF ≤ 6 |
| | more than 6: SC.DIFF > 6 |

probability of scoring. The first splits are made according to the variable SHOT.TYPE. After that, as expected on the basis of the preliminary analyses, SHOT.CLOCK has the most prominent role, immediately followed by MISS.PL for free throws and by covariates SC.DIFF, TIME, and POSS.TYPE, which seem to play a roles in interactions with the other variables.

In detail, the tree reveals the following relationships:

- For free throws, the only relevant variable is MISS.PL: the estimated scoring probability when the previous shot by that player scored a basket is 0.7481 vs. 0.7115 when the previous shot was missed.
- For 3-point shots, the first relevant variable is SHOT.CLOCK: the estimated scoring probability at time-end is 0.2939; otherwise, the scoring probability is 0.3504 and 0.3844 for middle-end and earlier shots, respectively, provided the game is not in quarter-end, when the scoring probability decreases to 0.3119.
- For 2-point shots, the most relevant variable is again SHOT.CLOCK: the estimated scoring probability is 0.4086, 0.4764, and 0.6580 for time-end, middle-end, and early shots, respectively; for early-middle shots, we distinguish between the game

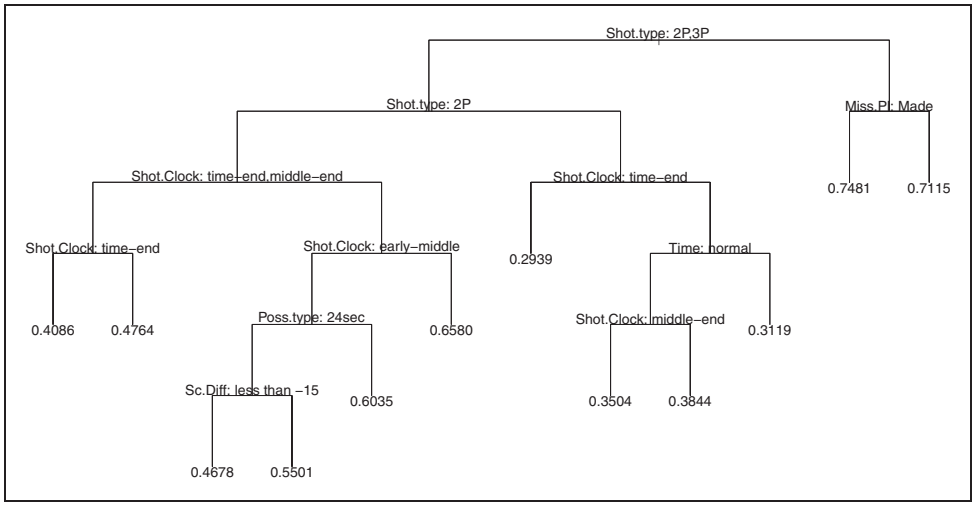


**Figure 11.** Classification tree for the dependent variable MADE with categorical covariates SHOT.CLOCK, SC.DIFF, MISS.T, Time, MISS.PL, SHOT.TYPE, POSS.TYPE, QUARTER (at each split, the declared splitting category refers to the left child node), dataset A2ITA.

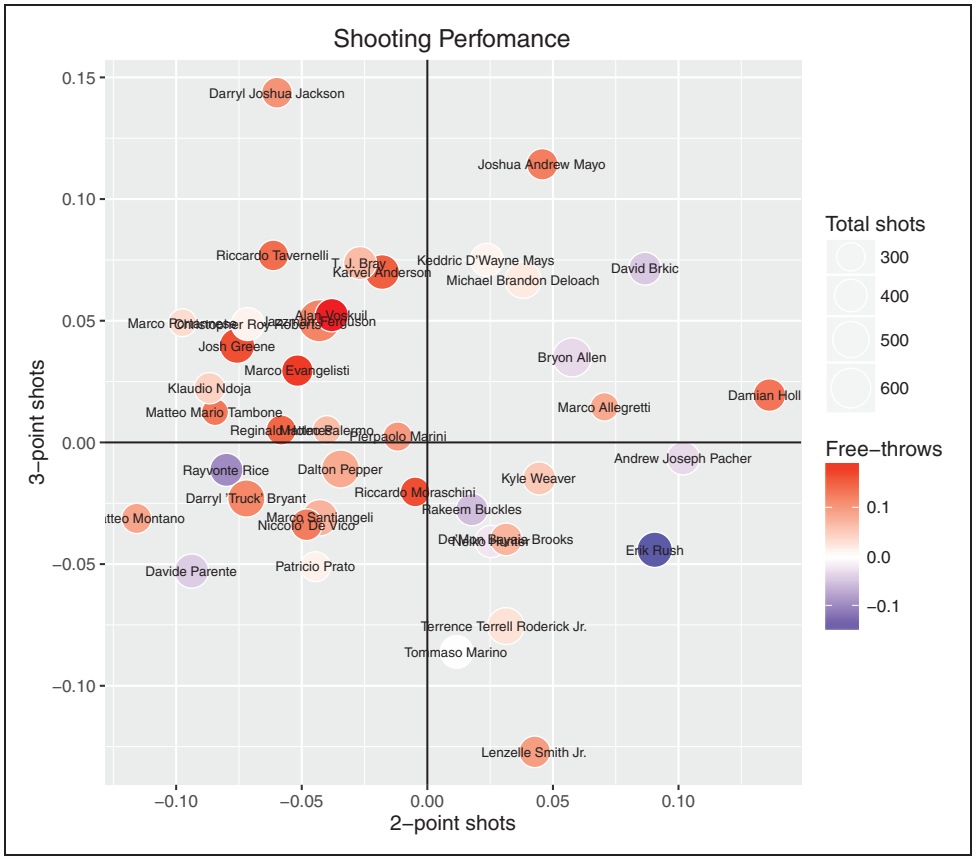


**Figure 12.** Players' shooting performances $P_i(2P)$ (*x*-axis), $P_i(3P)$ (*y*-axis), $P_i(FT)$ (blue-red scale), dataset A2ITA.

played after the shot clock has been reset to 14 s (i.e. in the first 4 s after the shot clock resetting, when the scoring probability increases to 0.6035) and 24-s possessions (a scoring probability of 0.4678 when the score difference is less than −15 and a probability of 0.5501 otherwise).

On the whole, the tree describes the impact of high-pressure game situations on the scoring probability from a multivariate perspective, and it highlights interesting interactions between different game situations. In particular, the tree reveals three interesting high-pressure situations: (a) in the last 2 s of the possession

(SHOT.CLOCK≤ 2); (b) when the score difference with respect to the opponent is low (−5 < SC.DIFF ≤ 1); (c) in the last 5 min of the match, when the score difference with respect to the opponent is in the range [−5;1]. The latter is similar to what in basketball analytics (see, for example, the NBA.coms stats page) is called "clutch time," which is defined as the last 5 min of any game in which the two teams are separated by 5 points or less. We decided to keep the definition in (c) obtained from the tree, according to a wider meaning of this term, which (together with "crunch time") is also used to indicate a period of intense pressure late in a game, with a small score difference. As pointed out in the Introduction section

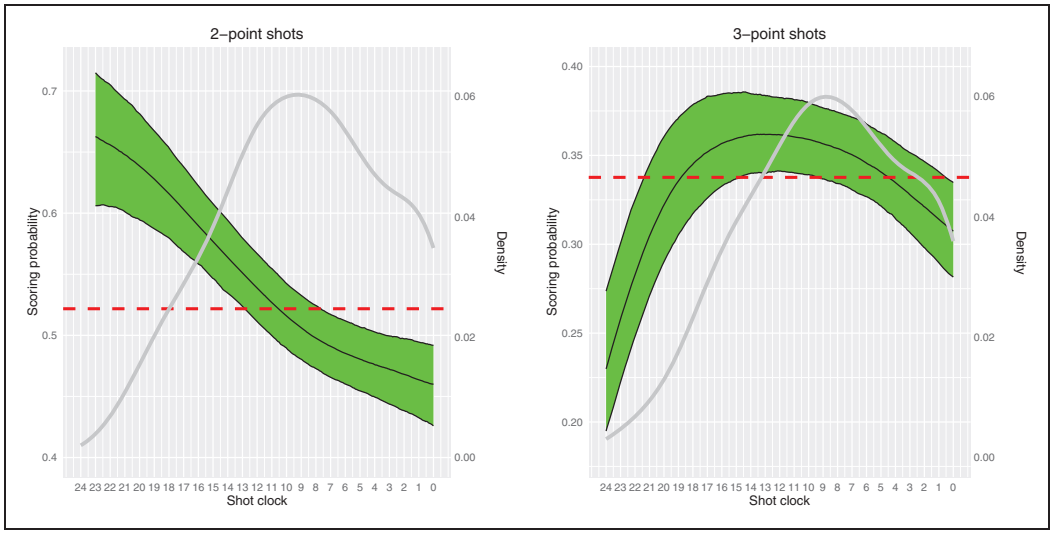


**Figure 13.** Bootstrap kernel estimates of the scoring probability conditional to SHOT.CLOCK (coloured areas) and distribution of shots by SHOT.CLOCK (gray lines), dataset RIO16.

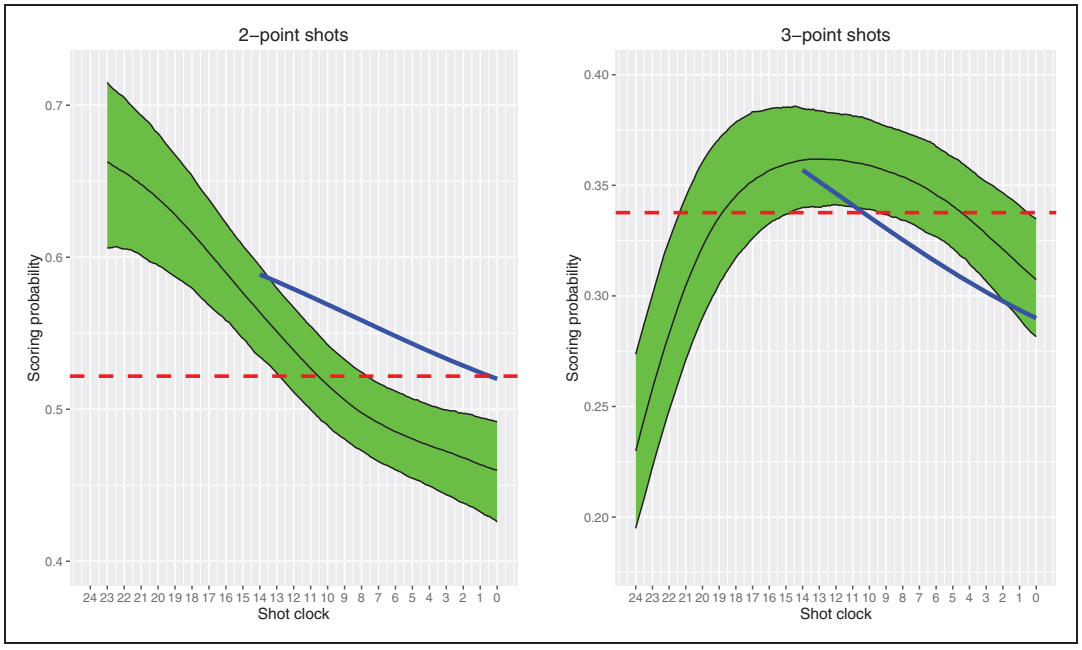


**Figure 14.** Kernel estimates of the scoring probability conditional to SHOT.CLOCK, with separate analysis of 14-s possessions (blue lines), dataset RIO16.

and highlighted by the preliminary analyses, high-pressure game situations are often managed differently by different players. For this reason, in Section ''Players' personal reactions to selected high-pressure game situations'', we will complete this analysis by investigating the players' personal reactions to the above-selected three high-pressure game situations.

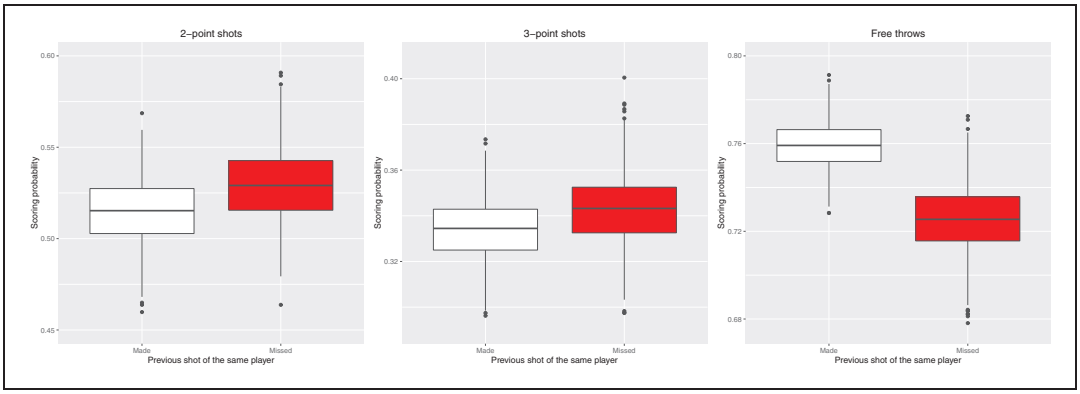


**Figure 15.** Boxplots of the bootstrap estimates of the scoring probability conditional to MISS.PL (white: previous shot made, red: previous shot missed), dataset RIO16.

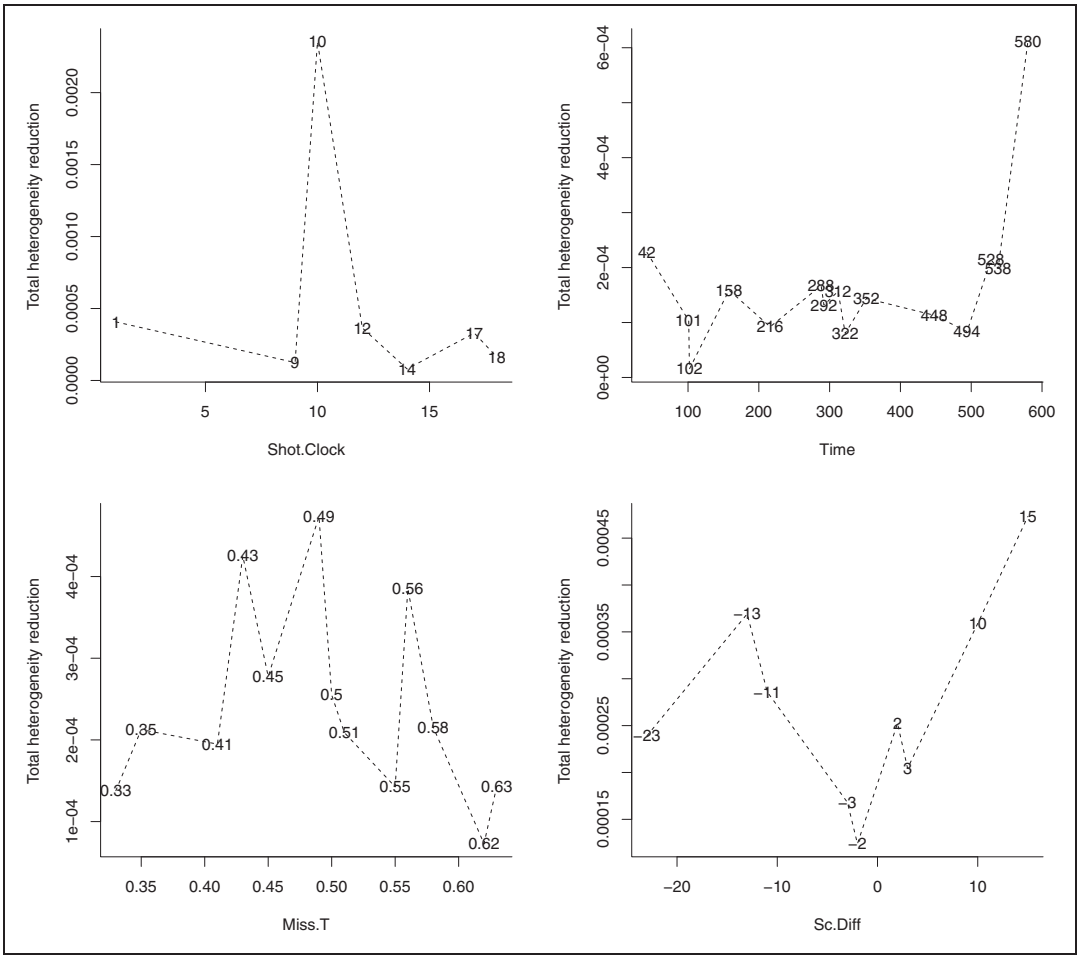


**Figure 16.** Thresholds importance measures, dataset RIO16.

## Shooting performance measures

The model presented above can now be used to develop new shooting performance measures, taking into account the circumstances under which each shot has been attempted. Our rationale is that, according to the evidence identified by the CART models, shots are not all alike. For example, a 2-point shot attempted in the last 2 s of the shot clock has a scoring probability of

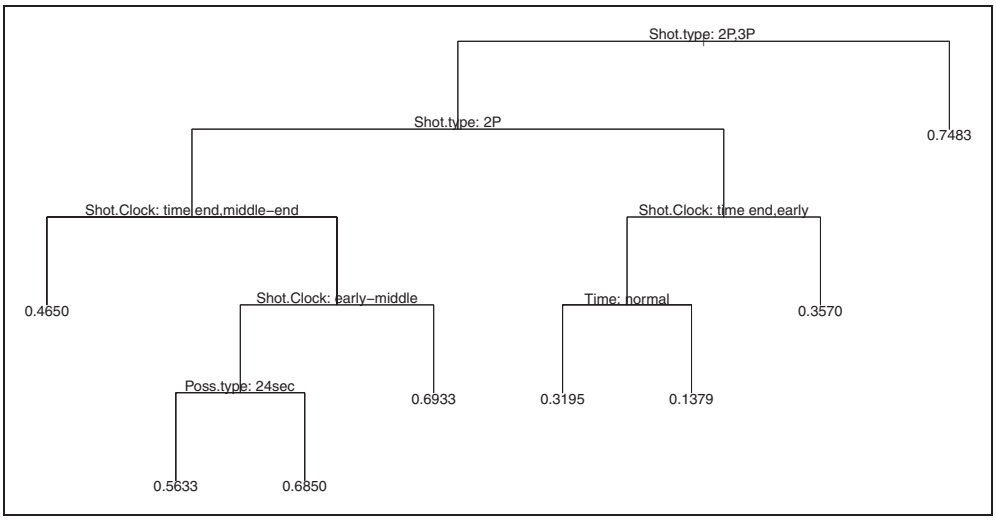


**Figure 17.** Classification tree for the dependent variable MADE with categorical covariates SHOT.CLOCK, SC.DIFF, MISS.T, Time, MISS.PL, SHOT.TYPE, POSS.TYPE, QUARTER (at each split, the declared splitting category refers to the left child node), dataset RIO16.

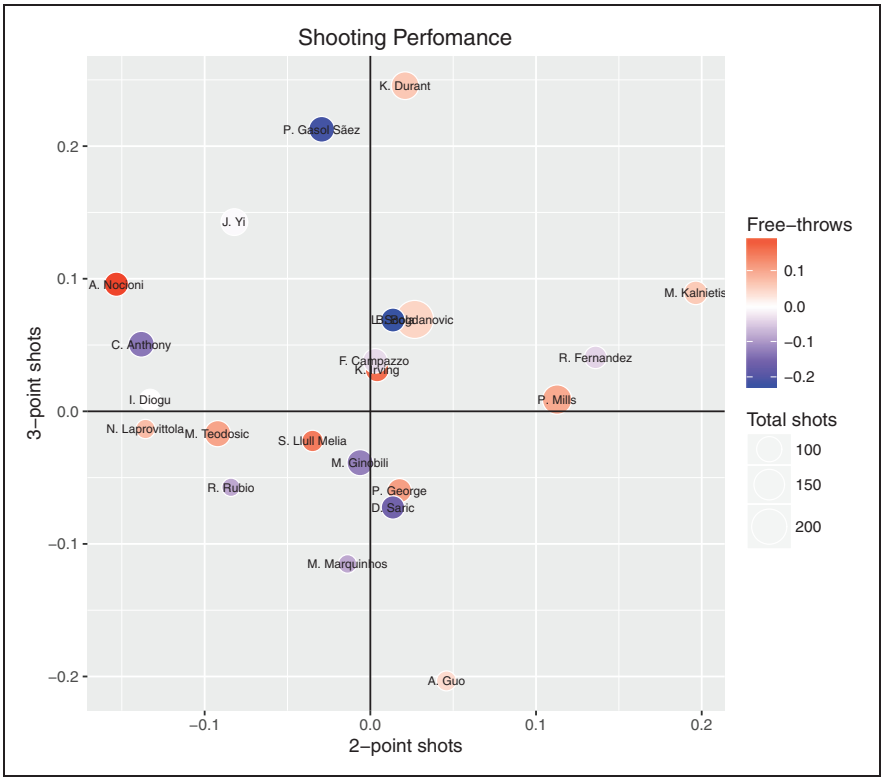


**Figure 18.** Players' shooting performances $P_i(2P)$ ($x$-axis), $P_i(3P)$ ($y$-axis), $P_i(FT)$ (blue-red scale), dataset RIO16.

approximately 40%, unlike a shot attempted in the first 7 s, which has a scoring probability greater than 65%. Our idea is to build a performance measure giving the right merit to each goal, according to the scoring probability of the corresponding shot, as opposed to the usual goal percentages, where each shot is considered to be the same as any other. For each shot type $T$ (2P: 2-point, 3P: 3-point, FT: free-throw), let $J_T$ be the set of attempted shots of type $T$. We denote with $x_{ij}$ the indicator assuming value 1 if the $j$th shot of the $i$th player scored a basket and 0 otherwise and with $\pi_{ij}$ its scoring probability according to the CART model; $\pi_{ij}$ is the scoring probability assigned by the CART model to the $j$th shot of the $i$th player, that is to a shot of the same type and attempted in the same game situation as the $j$th shot of the $i$th player. For each shot, the difference $x_{ij} - \pi_{ij}$ can be used as a performance measure of the shot. In fact, the difference is positive if the shot scored a basket (and the lower the scoring probability, the higher its value) and negative if it missed (and the higher the scoring probability, the higher its absolute value). For example, a basket is worth more when the scoring probability of the corresponding shot is low, whereas when a miss occurs, it is considered more detrimental when the scoring probability of the corresponding shot is high.

Then, we define the shooting performance of player $i$ for shot type $T$ as

$$P_i(T) = av_{j \in J_T}\left(x_{ij} - \pi_{ij}\right) \tag{1}$$

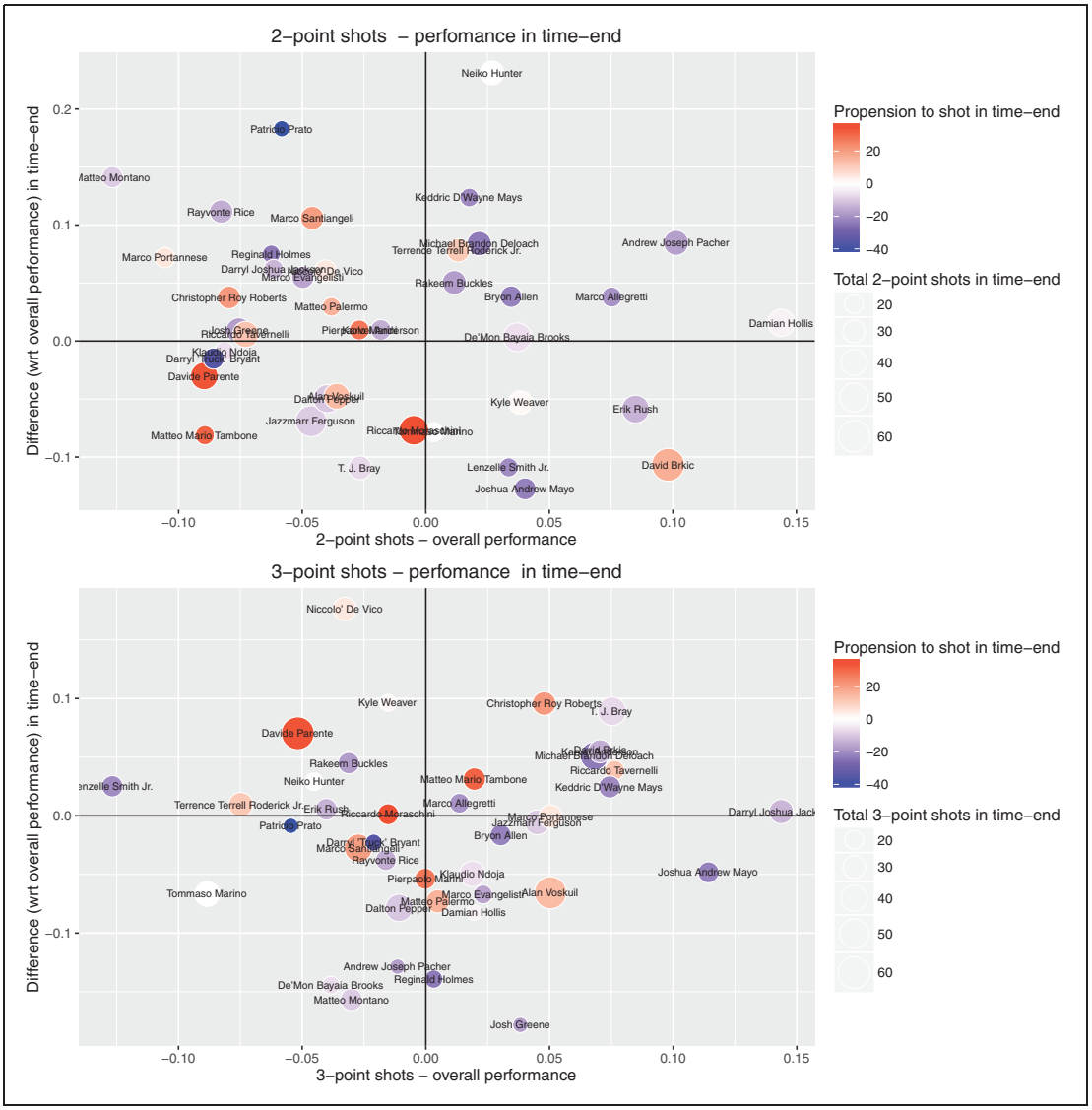


**Figure 19.** Players' personal reaction in the last 2 s of the possession, dataset A2ITA.

where $av_{j \in J_T}(\cdot)$ denotes averaging over all of the shots of type $T$ attempted by player $i$. The shooting performance proposed in equation (1) is obtained for each shot type as the average of the differences $(x_{ij} - \pi_{ij})$ that allow us to compare the performance of player $i$ with respect to $\pi_{ij}$, which can be viewed as the success probability of an "average" player (for that shot type in that game situation). Indeed, $\pi_{ij}$ is estimated in the CART leaf where the $j$th shot of the $i$th player falls as the proportion of made shots out of the total number of

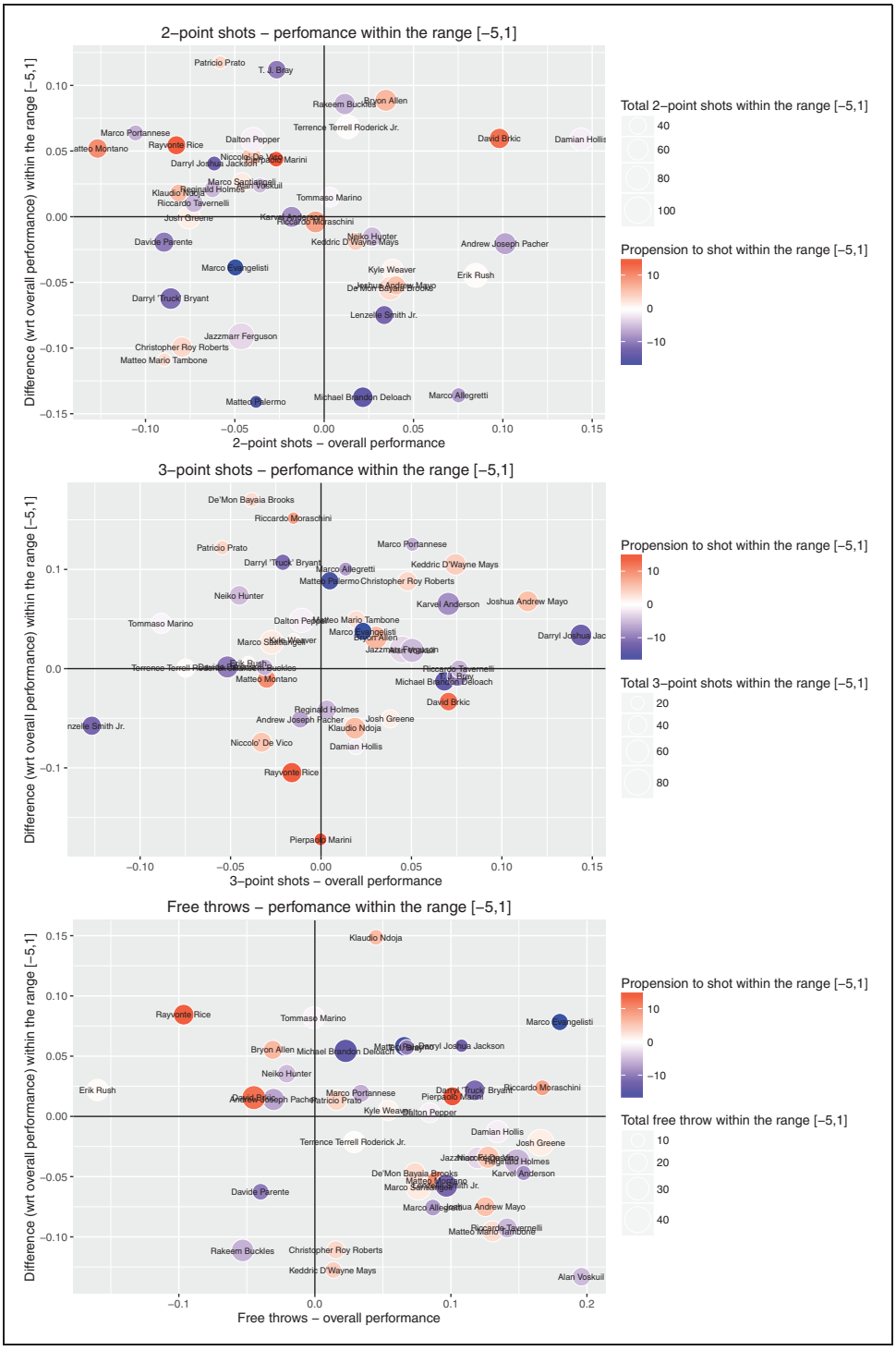


**Figure 20.** Players' personal reaction when score difference is within the range [−5;1], dataset A2ITA.

shots – of that type and in that game situation – attempted by all the players in the dataset. The value of $P_i$ can be interpreted depending on whether it is positive (meaning a positive balance between made and missed shots, taking into account their scoring probabilities) or negative and considering its absolute value. The highest interpretability is reached when players are compared by means of a graph, as in Figure 12, where each player is represented as a bubble with the size proportional to the total number

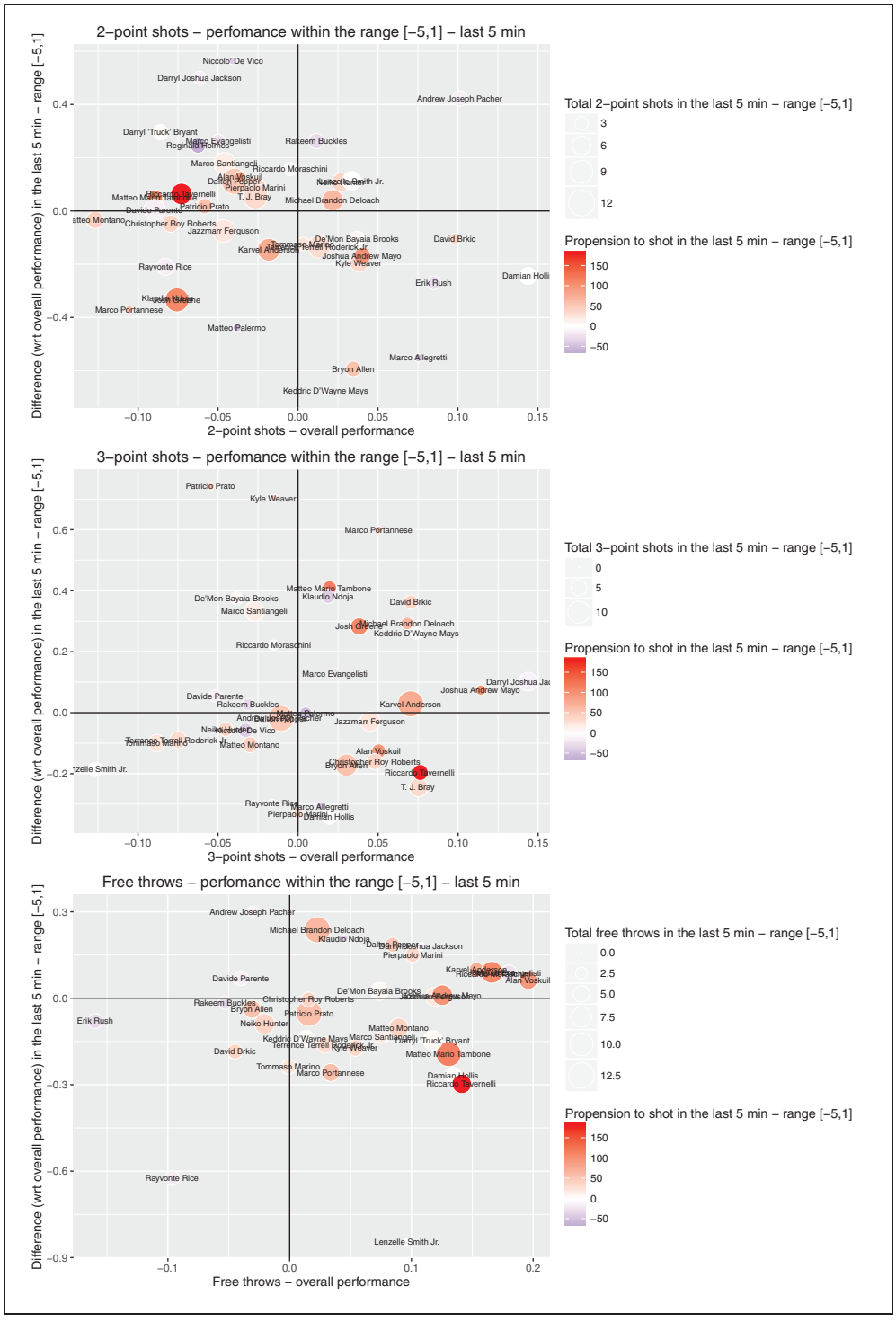


**Figure 21.** Players' personal reaction in last 5 min of the match when score difference is within the range [−5;1], dataset A2ITA.

of attempted shots. Each bubble is located in the plane with $P_i(2P)$ and $P_i(3P)$ on the $x$-axis and $y$-axis, respectively. The color represents the measure $P_i(FT)$, according to the blue–red scale reported in the graph. For clarity, the graphs display only players who attempted at least 70 shots for each shot type. Figure 12 shows which players are better (or worse) than the league average player, for each shot type and taking into account the particular game situation. For example, Erik Rush performs better than the average on 2-point shots but worse on 3-point shots and free throws. Most importantly, Figure 12 allows us to see some difference among players who, without considering the different game situations, have the same performance. For example, Bryon Allen and Marco Allegretti have the same 2-point field goal percentages (58%, computed on the entire 2015/2016 championship); however, if we consider the difficulty of the game situations in which they attempted their shots, Allegretti performs slightly better than Allen. Figure 12 also shows some players with different field goal percentages who receive equal evaluations when the difficulty of the game situation enters the performance measure. For example, Joshua Andrew Mayo and Kylie Waver receive a very similar performance measure, although they have different 2-point field goal percentages (54% and 57%, respectively).

It is interesting to understand the meaning of $P_i(T) = 0$, and what the origin of the axes in the graph represents. $P_i(T)$ is null when the $i$th player performs exactly according to the scoring probability, $\pi_{ij}$ (i.e. the success probability of the average player for that shot type in that game situation). In other words, let us consider all of the $N_k$ shots belonging to a certain type $T$, as determined by the $k$th CART leaf, where the estimated scoring probability is $\pi_k$. $P_i(T) = 0$ if the number of shots that scored a basket is exactly $N_k\pi_k$, and this holds for every leaf in the CART related to shots of type $T$. For example, based on the CART results in Figure 5, the estimated scoring probability for free throws is 0.7481 and 0.7115, according to whether the previous shot scored a basket or was missed, respectively. Let us assume that the $i$th player attempted 50 free throws when the previous shot scored a basket and 30 free throws when the previous shot was missed. Then, $P_i(FT) = 0$ in the theoretical circumstance that the $i$th player scored exactly $50 \times 0.7481$ goals when the previous shot scored a basket and $30 \times 0.7115$ goals when the previous shot was missed.

### Comparison to Rio16 data

As mentioned above, several studies claim that a player's reactions to high-pressure game situations may be different according to the professional level of the competition.[31] For this reason, we now check the stability of the most significant results obtained with the A2Ita dataset using the dataset Rio16, which comes from a profoundly different professional level, the 2016 Olympic Tournament. In detail, we check the relationships detected between the scoring probability and the variables Shot.Clock and Miss.Pl, as well as the multivariate associations highlighted by the CART.

No remarkable differences between datasets A2Ita and Rio16 are detected in terms of the effect of the shot clock on the scoring probability, except for a more pronounced effect on the probability of early 3-point shots and a reduced tendency to shoot in the very last seconds emerging from the Rio16 data (Figure 13 compared to Figure 1). The results are also quite similar with respect to the separate analysis of game play after the shot clock has been reset to 14 s (Figure 14 compared to Figure 2). The same can be said for the effect of the player's previous shot, where we find exactly the same pattern identified in the first analysis (Figure 15 compared to Figure 8). Also, in regard to the CART, it is noticeable that the main relationships detected for Rio16 are very similar to those exhibited by the A2Ita data, from the point of view of both the thresholds and the tree structure (Figures 16 and 17 compared to Figures 10 and 11). Here, due to the lower sample size, the tree is less robust, with leaf-node sizes ranging from 116 to 1,873. The AUC is 0.6822 (with 95% confidence interval [0.6696;0.6948])[42]. Finally, Figure 18 displays the bubble scatterplot describing the performances of players who attempted at least 15 shots of each shot type in dataset Rio16.

## Players' personal reactions to selected high-pressure game situations (Step 2)

In this section, we analyze players' personal reactions to the following three selected high-pressure situations (resulting from the tree in Subsection "Multivariate analysis with Classification and Regression Trees"):

(a) in the last 2 s of the possession (Shot.Clock $\leq 2$);
(b) when the score difference with respect to the opponent is in the range $[-5;1]$ (rigorously, $-5 <$ Sc.Diff $\leq 1$);
(c) in the last 5 min of the match, when the score difference with respect to the opponent is in the range $[-5;1]$

Recall that $x_{ij}$ denotes the indicator assuming value 1 if the $j$th shot of the $i$th player scored a basket and 0 if it did not, and let $J_S$ be the set of shots attempted in the selected high-pressure situation $S$. We introduce

the following statistics to be computed for each shot type $T$:

- Performance difference (with respect to the player's overall performance) in $S$, by shot type

$$D_i^{(S)}(T) = av_{j\in J_T\cap J_S}(x_{ij}-\pi_{ij}) - av_{j\in J_T}(x_{ij}-\pi_{ij}) = P_i^{(S)}(T) - P_i(T) \quad (2)$$

where $av_{j\in J_T\cap J_S}(\cdot)$ denotes averaging over all of the shots of type $T$ attempted by player $i$ in $S$. This statistic indicates whether the player's performance improves or gets worse in $S$ by computing the absolute increment or decrement in his performance measure in equation (1) when the shot is attempted during $S$.

- Propensity to shoot in $S$:
- for Case a, we compute the ratio between the fraction of shots attempted in the last 2 s of possession by player $i$ and the same fraction for the whole team

$$R_i^{(S)} = \left[ \frac{\frac{\sum_{j\in J_S} x_{ij}}{\sum_j x_{ij}}}{\frac{\sum_k \sum_{j\in J_S} x_{kj}}{\sum_k \sum_j x_{kj}}} - 1 \right] \cdot 100 \quad (3)$$

where the summation in the denominator is extended to all the players $k$ belonging to the same team as player $i$.

- for Cases b and c, we compute the ratio between the fraction of shots attempted in $S$ and the fraction of minutes played in $S$ by player $i$

$$R_i^{(S)} = \left[ \frac{\frac{\sum_{j\in J_S} x_{ij}}{\sum_j x_{ij}}}{\frac{t_i^{(S)}}{t_i}} - 1 \right] \cdot 100 \quad (4)$$

where $t_i$ and $t_i^{(S)}$ denote the minutes played by player $i$ in total and in situation $S$, respectively.

These statistics inform us on the extent to which the player is likely to risk taking a shot during $S$.

We can compute these statistics only when the sample size is large enough to guarantee a reasonable amount of observations, both in total and during $S$, for each player and for each shot type. For this reason, we will carry out the analysis only for the players who have attempted at least 70 shots for each shot type. For each situation $S$ and shot type $T$, we propose a graphical representation (Figures 19 to 21) where players are located in a plane with the performance index $P_i$ and the statistic $D_i$ on the $x$-axis and $y$-axis, respectively. Each player is represented by a bubble with size proportional to the number of shots he attempted in situation $S$ and color intensity proportional to the value of statistic $R_i$.

Using these graphs, coaches and scouts can analyze the personal reactions of each player to all of the selected situations for each shot type. For example, let us analyze the performance of Bryon Allen, member of the team "Mec-Energy Roseto" and winner of the MVP award among non-Italian players in the "Serie A2" Championship 2015/2016. With a huge number of total attempted shots (607 in the regular season), he exhibits very good performance for 2-point and 3-point shots, but he performs worse than the average for free throws (Figure 12). He shoots in the last 2 s of possession more rarely than the average for his entire team. In this situation, his performance tends to improve for 2-point shots and to (slightly) worsen for 3-point shots (Figure 19). When the score difference is within the range [–5,1] (Figure 20), his performance, with a high propensity to shoot, improves for all kinds of shots. Finally, in the last 5 min, when the score difference is within the range [−5;1], his propensity to shoot is moderate, but his performance worsens globally for all types of shots (Figure 21).

## Concluding remarks

In this paper, we analyzed players' shooting performance, taking account of factors that may generate high-pressure game situations. Considering play-by-play data from the Italian "Serie A2" Championship, we first carried out some exploratory univariate analyses, which allowed us to preliminarily identify factors affecting the scoring probability. Secondly, we investigated the presence of multivariate relationships and associations among game situations by means of a CART model, and we developed a new performance index based on it. The tree structure of the CART provided important insights into game mechanisms. In detail, it showed that the situation most impacting the scoring probability is when the shot clock is about to expire and, for free throws, when the player has missed the previous shot. In addition, the last 100 s of each quarter, the score difference, and the shot clock resetting to 14 s also play roles, in interaction with one another. The most important results obtained in this step were then validated using a dataset from the Olympic Basketball Tournament "Rio 2016", and we found

that the relationships continue to hold for players from a completely different professional level.

In a second step, we investigated the players' personal reactions to some selected high-pressure game situations by introducing some new measures, thus improving the indices currently used to measure shooting performance.

The proposed statistical methods and their application to a large amount of real data allowed us to obtain interesting results, as illustrated by some insightful graphical representations that can be exploited with the help of basketball experts to understand important behaviors of both teams and individual players. The analysis could be improved by the availability of additional data describing shots, such as the distance from the basket and the distance from the closest defender. In addition, this study could be extended by examining the physical skills and psychological characteristics of the players and then performing a cluster analysis to identify groups of players with similar profiles based on these traits.

## Acknowledgements

The authors thank the reviewers for their valuable comments, which greatly improved the paper and coach Marco Crespi for many stimulating discussions. The usual disclaimer applies.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research was carried out in collaboration with the Big&Open Data Innovation Laboratory (BODaI-Lab), University of Brescia (project nr. 03-2016, title ''Big Data Analytics in Sport'', bodai.unibs.it/BDSports).

## References

1. Albert J, Bennett J and Cochran JJ. *Anthology of statistics in sports* (vol. 16). Philadelphia, PA, USA: SIAM, 2005.
2. Albert J and Koning RH. *Statistical thinking in sports*. Boca Raton, FL, USA: CRC Press, 2007.
3. Zuccolotto P, Manisera M and Kenett RS (eds) Statistics in Sports. *Electronic Journal of Statistical Analysis* [Special issue] 2017; 10(3): 629–880.
4. Kubatko J, Oliver D, Pelton K, et al. A starting point for analyzing basketball statistics. *J Quant Anal Sports* 2007; 3: 1–22.
5. West BT, et al. A simple and flexible rating method for predicting success in the NCAA basketball tournament: updated results from 2007. *J Quant Anal Sports* 2008; 4: 1–18.
6. Loeffelholz B, Bednar E, Bauer KW, et al. Predicting NBA games using neural networks. *J Quant Anal Sports* 2009; 5: 1–15.
7. Brown M, Sokol J, et al. An improved LRMC method for NCAA basketball prediction. *J Quant Anal Sports* 2010; 6: 1–23.
8. Gupta AA. A new approach to bracket prediction in the NCAA men's basketball tournament based on a dual-proportion likelihood. *J Quant Anal Sports* 2015; 11: 53–67.
9. Lopez MJ and Matthews GJ. Building an NCAA men's basketball predictive model and quantifying its success. *J Quant Anal Sports* 2015; 11: 5–12.
10. Ruiz FJ and Perez-Cruz F. A generative model for predicting outcomes in college basketball. *J Quant Anal Sports* 2015; 11: 39–52.
11. Yuan LH, Liu A, Yeh A, et al. A mixture-of-modelers approach to forecasting NCAA tournament outcomes. *J Quant Anal Sports* 2015; 11: 13–27.
12. Manner H. Modeling and forecasting the outcomes of NBA basketball games. *J Quant Anal Sports* 2016; 12: 31–41.
13. Page GL, Fellingham GW and Reese CS. Using box-scores to determine a position's contribution to winning basketball games. *J Quant Anal Sports* 2007; 3: 1–16.
14. Cooper WW, Ruiz JL and Sirvent I. Selecting non-zero weights to evaluate effectiveness of basketball players with DEA. *Eur J Oper Res* 2009; 195: 563–574.
15. Piette J, Anand S and Zhang K. Scoring and shooting abilities of NBA players. *J Quant Anal Sports* 2010; 6: 1–23.
16. Fearnhead P and Taylor BM. On estimating the ability of NBA players. *J Quant Anal Sports* 2011; 7: 1–18.
17. Ozmen MU. Foreign player quota, experience and efficiency of basketball players. *J Quant Anal Sports* 2012; 8: 1–18.
18. Page GL, Barney BJ and McGuire AT. Effect of position, usage rate, and per game minutes played on NBA player production curves. *J Quant Anal Sports* 2013; 9: 337–345.
19. Deshpande SK and Jensen ST. Estimating an NBA player's impact on his team's chances of winning. *J Quant Anal Sports* 2016; 12: 51–72.
20. Skinner B. The price of anarchy in basketball. *J Quant Anal Sports* 2010; 6: 1–18.
21. Shortridge A, Goldsberry K and Adams M. Creating space to shoot: quantifying spatial relative field goal efficiency in basketball. *J Quant Anal Sports* 2014; 10: 303–313.
22. Annis DH, et al. Optimal end-game strategy in basketball. *J Quant Anal Sports* 2006; 2: 1–11.
23. Hand DJ, Mannila H and Smyth P. *Principles of data mining (adaptive computation and machine learning)*. Cambridge, MA, USA: MIT Press, 2001.
24. Han J, Pei J and Kamber M. *Data mining: concepts and techniques*. Waltham, MA, USA: Elsevier, 2011.
25. Tango TM, Lichtman MG and Dolphin AE. *The book: playing the percentages in baseball*. Washington DC, USA: Potomac Books, Inc., 2007.

26. Jones G. What is this thing called mental toughness? An investigation of elite sport performers. *J Appl Sport Psychol* 2002; 14: 205–218.
27. Apesteguia J and Palacios-Huerta I. Psychological pressure in competitive environments: evidence from a randomized natural experiment. *Am Econ Rev* 2010; 100: 2548–2564.
28. Goldman M and Rao JM. Effort vs. concentration: the asymmetric impact of pressure on NBA performance. In: *MIT Sloan sports analytics conference*, Boston, MA, USA, 4–5 March 2012.
29. Taylor J. Predicting athletic performance with self-confidence and somatic and cognitive anxiety as a function of motor and physiological requirements in six sports. *J Pers* 1987; 55: 139–153.
30. Madden CC, Kirkby RJ, McDonald D, et al. Stressful situations in competitive basketball. *Austr Psychol* 1995; 30: 119–124.
31. Madden C, Summers J, Brown D, et al. The influence of perceived stress on coping with competitive basketball. *Int J Sport Psychol* 1990; 21: 21–35.
32. Trninić S, Dizdar D and Lukšić E. Differences between winning and defeated top quality basketball teams in final tournaments of European club championship. *Coll Antropol* 2002; 26: 521–531.
33. Carpita M, Sandri M, Simonetto A, et al. Football mining with R [Chapter 14]. In: Zhao Y and Cen Y (eds) *Data mining applications with R*. Cambridge, MA, USA: Elsevier, 2014. pp.397–433.
34. Carpita M, Sandri M, Simonetto A, et al. Discovering the drivers of football match outcomes with data mining. *Qual Technol Quant Manage* 2015; 12: 561–577.
35. Breiman L, Friedman J, Stone CJ, et al. *Classification and regression trees*. Belmont, CA, USA: Wadsworth, CRC Press, 1984.
36. Wu X, Kumar V, Quinlan JR, et al. Top 10 algorithms in data mining. *Knowl Inform Syst* 2008; 14: 1–37.
37. Breiman L. Bagging predictors. *Mach Learn* 1996; 24: 123–140.
38. Breiman L. Random forests. *Mach Learn* 2001; 45: 5–32.
39. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001; 29: 1189–1232.
40. Sandri M and Zuccolotto P. A bias correction algorithm for the Gini variable importance measure in classification trees. *J Comput Graph Stat* 2008; 17: 611–628.
41. Sandri M and Zuccolotto P. Analysis and correction of bias in total decrease in node impurity measures for tree-based algorithms. *Stat Comput* 2010; 20: 393–407.
42. Robin X, Turck N, Hainard A, et al. p ROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform* 2011; 12: 77.