THE GENERALIZED SHAPLEY MEASURE FOR RANKING PLAYERS IN BASKETBALL: APPLICATIONS AND FUTURE DIRECTIONS

Rodolfo Metulini[1], Francesco Biancalani[2] and Giorgio Gnecco[2]

[1] Department of Economics, University of Bergamo (e-mail: `rodolfo.metulini@unibg.it`)

[2] Laboratory for the Analysis of CompleX Economic Systems (AXES), IMT School for Advanced Studies (e-mail: `francesco.biancalani@imtlucca.it`, `giorgio.gnecco@imtlucca.it`)

ABSTRACT: A wide range of measures have been proposed to quantify a player's marginal contribution to a team. We contributed to this strand of research by proposing, specifically for basketball, a new measure based on a combination of the Shapley value from game theory and the logistic regression, which is based on considering the utility of a player in every single lineup. Some applications where the measure can be useful are presented, such as ranking players, forming lineups, and predicting a remunerative new contract for free agent players. We also discuss possible ideas for future research developments.

KEYWORDS: Sports statistics, Shapley value, logistic regression, players ranking, statistical learning.

1 Introduction & state of the art

Thanks to advancements in technologies and the related increase of available data, measuring the importance of players in team sports to help coaches and staff to win more games is gaining relevance. A wide collection of synthetic indices has been developed in the sport statistics literature to measure each player's contribution to the team win. Among others, we can mention Plus-Minus (PM) and its generalizations (see, e.g., Kubatko *et al.*, 2007; Grassetti *et al.*, 2021), Win-Shares (WS), Wins Above Replacement Player (WARP) and their advances (see, for a review, Sarlis & Tjortjis, 2020, which also highlights pros and cons of such methods). A new measure of players' contribution to the team in basketball has been recently developed (Metulini & Gnecco, 2022). It adopts a combination of a two-step approach based on the logistic regression and the concept of generalized Shapley value (Nowak & Radzik, 1994). This proposal aims to gather most of the advantages (and avoid disadvantages) of industry-standard measures. Recent PM versions moved in the direction of solving some cons, such as just considering only scoring factors, and multicollinearity. However, those issues still need attention (Terner & Franks, 2021). The measure proposed in Metulini & Gnecco, 2022, similarly to BPM, presents the advantage of being based on both offensive and defensive scoring and non scoring features. Furthermore, the method takes into account probabilities to win the game, which are estimated based on a long time span of box-score synthetic measures (the so-called four Dean's factors, Kubatko *et al.*, 2007) that produce extremely high goodness of fit. Moreover, similarly to what WARP does by introducing the replacement level player, the approach proposed in Metulini & Gnecco, 2022, considering lineups, accounts for marginal utilities of players. This is achieved by explicitly accounting for all the lineups each player has played with. In doing so, considering a proper level for the replacement player is not needed and multicollinearity is avoided.

2 The generalized Shapley measure

The generalized Shapley value for a player in a generalized coalitional game with n players represents his/her average marginal utility to a suitably randomly formed ordered coalition of players. To obtain this measure for basketball players, first, the coefficients of a logistic model applied to game level are computed through the equation $log\frac{P(Y_i=1|\boldsymbol{X})}{P(Y_i=0|\boldsymbol{X})} = \boldsymbol{X}_i\boldsymbol{\beta}$, where the left part of the equation represents the log-odd of Y_i conditional on $\boldsymbol{X}$; $\boldsymbol{Y}$ is the binary response variable representing the outcome of the games, $Y_i \in \{0,1\}$, $i = 1,...,g$, where g is the number of games. $\boldsymbol{X}_i$ is the $i-$th row of the design matrix $\boldsymbol{X}$ with g rows and p columns (p=8, the eight Dean's factors used as explanatory variables computed at the game level). $\boldsymbol{\beta}$ is a vector containing the p regression parameters associated with the explanatory variables. Since the single lineup does not play the full match, to determine the probabilities to win the game for that quintet is not feasible. To deal with this issue, a dataset $\tilde{\boldsymbol{X}}$ where the Dean's factors are computed at the single lineup level (i.e., each row of the dataset corresponds to a lineup) is used and the probabilities to win the game $P(Win)_{L_j}$ is predicted on each lineup L_j by using the vector $\hat{\boldsymbol{\beta}}$ of estimated coefficients from the first step. Let $\tilde{\boldsymbol{X}}_{\boldsymbol{j}}$ be the j-th row of the matrix $\tilde{\boldsymbol{X}}$ with l rows (where l is the number of lineups considered) and p=8 columns (expressing the eight Dean's factors computed at the lineup level). The probabilities to win the game for the lineup L_j is expressed as $P(Win)_{L_j} = \frac{exp(\tilde{\boldsymbol{X}}_j\hat{\boldsymbol{\beta}})}{1+exp(\tilde{\boldsymbol{X}}_j\hat{\boldsymbol{\beta}})}, j = 1,...,l$. In the third step, one considers two versions (*unweighted* and *weighted*)[*] of the generalized characteristic function, hence of the (Nowak-Radzik, NR) generalized

---

[*] The two differ in terms of taking/not taking in account the time players are on the court.

Shapley value: $\phi_i^{NR}(N,\upsilon) = \frac{1}{n!}\sum_{T\in\mathcal{T}\,\text{with}\,|T|=n}(\upsilon((T(i),i)) - \upsilon(T(i)))$, where $\mathcal{T}$ refers to the set of all ordered coalitions of players, $T(i)$ represents the ordered subcoalition made by the predecessors of i in the permutation T, whereas $(T(i),i)$ is the ordered subcoalition made by $T(i)$ followed by i. $\upsilon : \mathcal{T} \to \mathbb{R}$ (such that $\upsilon(\emptyset) = 0$) is called generalized characteristic function. Metulini & Gnecco, 2022 described two possible choices for the generalized characteristic function $\upsilon(.)$ ($\upsilon_1(.)$ and $\upsilon_2(.)$). When restricted to a lineup, the generalized characteristic function $\upsilon_1(.)$ represents the probability $P(Win)$ to win the game for every specific lineup. At the same time, $\upsilon_2(.)$ is a function of both $P(Win)$ and the probability of occurrence $P(Occ)$ of that lineup on the court. The corresponding generalized Shapley value measures are called unweighted generalized Shapley value (UWGS) and generalized Shapley value (WGS).

3 Applications

The UWGS (WGS) may be used for different purposes. For example, Metulini & Gnecco, 2022, by computing the generalized characteristic function based on all the games (regular season and playoff) from 17 National Basketball Association (NBA) seasons (2004/2005 – 2020/2021)[†], determine (approximations of) the two measures for the Utah Jazz players during season 2020-21, rank players in terms of such measures, and propose a "greedy" algorithm to suggest best lineups conditional to the presence/absence of a specific team player. The algorithm is based on choosing the player with the largest UWGS (WGS), recomputing the UWGS (WGS) of teammates based just on the lineups where the chosen player was in, and repeating the process until five players have been chosen. Since the UWGS and the WGS are composite measures that aim to evaluate a player marginal utility in terms of winning the game, it is reasonable to think that a player may be rewarded with a salary that is proportional to these measures. Biancalani *et al.*, 2023 using income data available at `basketballinsiders.com` and computing such measures for the players of three NBA teams, proposed an instrument to predict the deal of a better contract (compared to the previous year) in the next season based on deviations of estimated salaries (according to a log-linear model) from the true incomes.

4 Possible developments

From a methodological viewpoint, a natural future direction might regard developing a generalized Shapley measure that takes into account players' roles

[†] Features of the logistic model's dependent variable and Dean's factors for both $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ are computed based on the dataset provided by BigDataBall Company (UK) (`www.bigdataball.com`).

as constraints. In fact, with the UWGS (WGS), we might obtain (potentially) that the players with the largest marginal utility are all playing the same role. However, when using the UWGS (WGS) to rank players, forming a lineup with five players of the same role does not make sense. A solution to this issue might be that of classifying players in the same role (by using a cluster analysis), then compute the UWGS (WGS) separately for each role. From an applied point of view, players' popularity retrieved from Google Trends (`trends.google.it/home`) may be exploited to investigate the degrees of relationship between the player's marginal utility and his/her popularity.

Acknowledgements

The authors acknowledge partial support from the INdAM-GNAMPA 2023 project "Sviluppo di metodi di machine learning per la stima del valore Shapley e di sue generalizzazioni", code CUP_E53C22001930001.

References

BIANCALANI, FRANCESCO, GNECCO, GIORGIO, METULINI, RODOLFO, *et al.* 2023. the relationship between players' average marginal contributions and salaries: an application to NBA basketball using the generalized Shapley value. *Statistica Applicata*, 1–29.

GRASSETTI, LUCA, BELLIO, RUGGERO, DI GASPERO, LUCA, FONSECA, GIOVANNI, & VIDONI, PAOLO. 2021. An extended regularized adjusted plus-minus analysis for lineup management in basketball using play-by-play data. *IMA Journal of Management Mathematics*, **32**(4), 385–409.

KUBATKO, JUSTIN, OLIVER, DEAN, PELTON, KEVIN, & ROSENBAUM, DAN T. 2007. A starting point for analyzing basketball statistics. *Journal of quantitative analysis in sports*, **3**(3).

METULINI, RODOLFO, & GNECCO, GIORGIO. 2022. Measuring players' importance in basketball using the generalized Shapley value. *Annals of Operations Research*, 1–25.

NOWAK, ANDRZEJ S, & RADZIK, TADEUSZ. 1994. The Shapley value for *n*-person games in generalized characteristic function form. *Games and Economic Behavior*, **6**(1), 150–161.

SARLIS, VANGELIS, & TJORTJIS, CHRISTOS. 2020. Sports analytics—Evaluation of basketball players and team performance. *Information Systems*, **93**, 101562.

TERNER, ZACHARY, & FRANKS, ALEXANDER. 2021. Modeling player and team performance in basketball. *Annual Review of Statistics and Its Application*, **8**, 1–23.