Statistica Applicata - Italian Journal of Applied Statistics Vol. 35 (2)
doi.org/10.26398/IJAS.0035-009

A SURVIVAL ANALYSIS STUDY TO DISCOVER WHICH SKILLS DETERMINE A HIGHER SCORING IN BASKETBALL

Ambra Macis [1], **Marica Manisera, Paola Zuccolotto**
Department of Economics and Management, University of Brescia, Italy

Marco Sandri
Big and Open Data Innovation Laboratory, University of Brescia, Italy

Abstract *Over the years data analytics for sports has developed consistently. Survival analysis is a method that allows to study the occurrence of a particular event during a period of follow-up. This work aims studying the main achievements associated to the probability of reaching a certain amount of points during a NBA season segment. A Stepwise Cox regression model and a Lasso Cox regression were used to select the most important variables. Two settings were examined, with* 20% *and* 50% *censoring. Results showed that attempting more shots, gaining more achievements (double doubles) and having been selected for the All-Star game increase the probability of success, i.e. exceeding the given threshold of points. Moreover, a higher number of steals seems to decrease the probability of reaching a certain amount of points. Thus, players more involved in this fundamental are penalized in terms of scored points.*

Keywords: *Stepwise Cox regression; Lasso Cox regression; Basketball analytics; Performance.*

1. INTRODUCTION

Sport analytics has developed consistently over the years. For what concerns basketball, Data Science has been widely used to answer different questions and several studies have been carried out with a wide variety of aims. Just as an example, contributions in the literature deal with the analysis of players' performance and of the impact of high pressure game situations, the prediction of the outcomes of a game or a tournament, the identification of factors that distinguish successful and unsuccessful teams and the monitoring of playing patterns with reference to roles (Zuccolotto and Manisera, 2020).

In this work we deal with survival analysis, a class of statistical methods devoted to the study of the occurrence of an event during a given observation time. This kind of analysis was firstly introduced for applications in medicine

[1] Ambra Macis ambra.macis@unibs.it

for studying, for example, disease recurrence or death; however, it has also been widely used in sport analytics. Survival analysis has been used in the context of sports with several different aims, such as, for example, for studying the relationship between specific features and dropout of young athletes in many sports (Back et al., 2022; Moulds et al., 2020; Pion et al., 2015; Smith and Weir, 2022), or for evaluating the career length of professional basketball players (Fynn and Sonnenschein, 2012). Other studies analyzed the criterion determining the decision of a football coach of doing the first substitution during a match (Del Corral et al., 2008), the effect of team performance in the dismissal of coaches (Tozetto et al., 2019; Wangrow et al., 2018), the duration of Olympic success (Csurilla and Fertő, 2022; Gutiérrez et al., 2011), whether Olympic medalists live longer than the general population (Clarke et al., 2012). Furthermore, many studies dealt with injury prevention and the prediction of risk factors for injury (Beynnon et al., 2005; Buist et al., 2010; Ekeland et al., 2020; Hopkins et al., 2007; Lu et al., 2022; Mahmood et al., 2014; Venturelli et al., 2011; Zumeta-Olaskoaga et al., 2021) and recovery after injuries and sport-related concussions (Dekker et al., 2017; Howell et al., 2019; Jack et al., 2019; Kontos et al., 2019; Lawrence et al., 2018; Mai et al., 2017; Nelson et al., 2016; Sochacki et al., 2019). Finally, other works analyzed the impact of performance indicators on the time when the first goal is scored or the effect of that time on the following goal in football (Nevo and Ritov, 2013; Pratas et al., 2016) or studied times between goals in ice hockey (Thomas, 2007).

Up to now, to the best of our knowledge, survival analysis has not been used for studies in which the event of interest is a measure of the overall performance of a player. This work aims indeed to study the offensive performance of the National Basketball Association (NBA) players in a novel way, using survival analysis. In details, the player's performance has been measured in terms of exceeding of a given amount of points during a season segment, and the interest has been focused on the identification of the main achievements related to the occurrence of this event. Thus, from a statistical point of view this means performing a well-defined variable selection with only few variables selected. For this reason, the Lasso Cox has been chosen because it allows to select only few variables from all those taken into account. Moreover, the Stepwise Cox regression has been used as additional method to have a term of comparison.

The article is organized as follows. The following section reports the methodological framework. Then, Sections 3 and 4 show respectively the data used for carrying out the study and the obtained results. The paper ends with the final discussion.

2. METHODS

Survival analysis aims to study the occurrence of a particular event during an observed period of time. The main feature of this kind of data is censoring. A subject is censored when for him/her the event of interest has not been observed during the observation time, so that the only known thing is the last time he/she did not experience the event (Collett, 2015). In this context a subject is denoted by three elements: (i) a time point τ that can be the observed time t or the censoring time c; (ii) an event indicator δ that equals 1 if the subject experienced the event and 0 if he/she is censored; and (iii) a vector of observed covariates $\boldsymbol{x}$. More in detail, for the i^{th} subject:

$$\tau_i = \min(t_i, c_i) = \begin{cases} t_i & \text{if } \delta_i = 1 \\ c_i & \text{if } \delta_i = 0 \end{cases} . \tag{1}$$

The actual survival time can be seen as the observed value of a non-negative random variable T with density function $f(t)$, such that $f(t) \geq 0$ and $\int_0^{+\infty} f(t)dt = 1$. Key elements in survival analysis, which allow to specify the probability distribution of T, are the survival and hazard functions.

The survival function $S(t)$ measures the probability that an individual survives (does not experience the event) beyond a given timepoint t and can be defined as

$$S(t) = P(T \geq t) = \int_t^{\infty} f(u)du . \tag{2}$$

This function, that is the complementary to one of the cumulative distribution function $F(t)$, is non-increasing and right-continuous with $S(0) = 1$ and $\lim_{t \to +\infty} S(t) = 0$.

The hazard function measures the probability that an individual has the event of interest at time t conditional that the event has not occurred until that time. Formally, it is defined as

$$h(t) = \frac{f(t)}{S(t)} = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} . \tag{3}$$

It specifies the instantaneous rate at which events occur for subjects that are surviving at time t.

2.1. COX REGRESSION MODEL

The Cox proportional hazards (PH) regression model (Cox, 1972) is one of the most used classical methods for analyzing survival data. It allows to estimate the hazard of a subject depending on a set of covariates. The main assumption of the model is the proportionality of hazards, implying that the hazards of two groups of subjects, $h_1(\cdot)$ and $h_2(\cdot)$, are proportional, so that their ratio is constant over time:

$$\Psi = \frac{h_1(t)}{h_2(t)} \qquad \forall t, \tag{4}$$

where Ψ is a constant called *hazard ratio* (HR) or *relative hazard*.
The Cox PH model can then be expressed as

$$h_i(t) = h_0(t)\Psi = h_0(t)e^{\beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki}} = h_0(t)e^{\sum_{k=1}^{K} \beta_k x_{ki}} ,$$

where $h_i(t)$ represents the hazard function for the i^{th} subject; $h_0(t)$ is the baseline hazard, that is the risk for a subject whose values of all the independent variables are equal to zero; x_{ki} is the observed value of the k^{th} covariate for the i^{th} subject and β_k is the related coefficient.

Due to the presence of censoring, only a partial likelihood can be considered. The partial log-likelihood function can be expressed as

$$\ln L(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{\beta}' \boldsymbol{x}_i - \ln \left(\sum_{l \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_l} \right) \right] = \sum_{i=1}^{n} \delta_i \boldsymbol{\beta}' \boldsymbol{x}_i - \ln \left(\sum_{l \in R_j} e^{\boldsymbol{\beta}' \boldsymbol{x}_l} \right) \sum_{i=1}^{n} \delta_i . \tag{5}$$

where n is the number of subjects in the sample, $\boldsymbol{x}_i$ is the observed covariate vector for the i^{th} subject who experienced the event at the j^{th} ordered event time $t_{(j)}$ and R_j is the set of subjects at risk (risk set) at time $t_{(j)}$. According to this expression, only uncensored subjects ($\delta_i = 1$) have a direct effect on (5); on the contrary, censored observations do not directly contribute to the likelihood, but indirectly enter in the likelihood function because all the subjects are included in the risk set.
In presence of ties, approximations of the two functions are needed (Collett, 2015).

Then, the β coefficients are estimated maximizing the partial log-likelihood, using iterative methods as the Newton-Raphson algorithm (Collett, 2015).
Each coefficient represents the estimated change in the logarithm of the hazard ratio in correspondence of a change of the corresponding covariate. Usually, their

exponential is considered, measuring the hazard ratio. A value of e^{β_k} greater (lower) than 1 indicates that for a one-unit increase in the continuous variable X_k, the hazard increases (decreases) by e^{β_k}, or, in an analogous way, if X_k is categorical, that a subject in group k has a higher (lower) hazard (equal to e^{β_k}) relative to a subject in the reference group.

The Cox PH model is called semiparametric because it is based on a nonparametric component (the baseline hazard - no distributional assumption is made for survival times) and on a parametric term.

Once the model has been fitted, stepwise variable selection can be used, based on different information criteria, as the Akaike Information Criterion (AIC). This selection method can be an efficient way to select a parsimonious model, because of the limited computation time and the possibility of tracking the variable selection process easily, allowing also to have further information on the excluded variables. However, it has also some disadvantages. Among these, there are (i) multiple comparisons problems and (ii) biased regression coefficient estimates (Harrell, 2015). Moreover, the obtained results may depend on the criterion used and on the ordering of the selected variables. Furthermore, another disadvantage concerns the fact that it is not possible to carry out an exhaustive analysis of all the possible combinations of the K predictors. Finally, in many situations, variable selection using stepwise regression shows a high unstability, especially when the sample size is small compared to the number of candidate variables, because many variable combinations can fit the data in a similar way (Derksen and Keselman, 1992).

2.2. REGULARIZED COX REGRESSION MODEL

In high-dimensional data contexts, usually, the interest is on variable selection in order to identify, among the many available variables, the most important ones. Thus, variable selection helps determining all the (informative) variables that are strictly related to the outcome, removing uninformative variables that decrease the precision and increase the complexity of the model. So, variable selection provides a balance between parsimony and goodness of fit of the model.

To this extent, regularized models are a good choice because they can allow to obtain a sparse model with many estimated coefficients equal to zero. In particular, they are based on the minimization of a loss function under a constraint that penalizes the flexibility of the model. Also, for Cox PH regression model different regularizations have been proposed (e.g. Lasso, Ridge, Elastic Net). Among

these, the Lasso (Least Absolute Shrinkage and Selection Operation), firstly proposed by Tibshirani in 1997 (Tibshirani, 1997), is one of the most used if the aim is variable selection. This because the Lasso is based on the use of a ℓ_1-norm penalty, that allows to obtain a well-defined solution with few nonzero coefficients β_k (Simon et al., 2011). Therefore, the Lasso is advantageous in terms of interpretation of the model and computational convenience (Hastie et al., 2015). Moreover, it is an interesting and useful method because it simultaneously performs feature selection among all the covariates and estimates the regression coefficients. The only requirement is to have standardized variables. However, the Lasso has also some limitations related to the possibility of obtaining biased estimates; so, it is not possible, for example, to combine the estimates with standard errors and to make inference through the estimation of confidence intervals and hypothesis testing.

Regularized parameters can be obtained by minimizing the negative partial log-likelihood $l(\beta)$ (see Equation (5)) under the constraint that the sum of the absolute values of the parameters is bounded by a constant (the Lasso penalty):

$$\hat{\beta} = argmin_{\beta} - l(\beta) \text{ subject to } \sum_{k=1}^{K} |\beta_k| \leq s \text{ , with } s > 0.$$

The regularization parameter s is a non-negative tuning parameter that controls the impact of the penalty. The larger the value, the lower the amount of shrinkage (Ekman, 2017).

K-fold cross-validation is used for identifying the best parameter s; the optimal regularization parameter is the one that minimizes the cross-validation error. In survival analysis one of the most used performance measures is the Harrell's Concordance index (C-index), a ranking measure based on the concordance of observed and predicted values (Harrell et al., 1982). This index is therefore used for measuring the cross-validation error during the estimation of the regularized parameters. As a higher C-index value means a better performance, the cross-validation error is measured as $1 - C$ (Tay et al., 2022).

The Lasso technique for variable selection in the Cox model is a worthy competitor to stepwise selection (Tibshirani, 1997), a variable selection procedure usually performed on the basis of AIC. From the simulation studies performed by Tibshirani in 1997 (Tibshirani, 1997) it emerged that the Lasso is less variable than the stepwise Cox, still yielding interpretable models.

In the case study that will be presented in the next section, analyses have been performed by using *R* (R Core Team, 2021). In particular, `survival` and `glmnet` packages have been used for estimating the Cox and the Lasso Cox regression models. Finally, `riskRegression` has been used for evaluating the time-dependent Area Under the Curves (AUCs).

3. DATA AND STUDY DESIGN

NBA data were analyzed. In particular, we considered the 2020-2021 regular season and divided it in two segments, the pre- and the post- All-Star (AS) game. The pre-AS game data have been used for extracting the baseline covariates (retrieved from the NBA website), while, play-by-play data have been used for the follow-up. This dataset has been kindly made available by BigDataBall (`www.bigdataball.com`), a reliable source of validated and verified data for the NBA, the Major League Baseball (MLB), the National Football League (NFL) and the Women's NBA (WNBA). All the variables included in this study have been chosen to characterize the performance abilities of each player.
A sample of $n = 359$ players has been considered. For each player a set of $K = 34$ baseline covariates $\boldsymbol{X} = (X_1, X_2, ..., X_K)$ corresponding to the main achievements gained in the pre-AS game has been observed. Let's denote with $\boldsymbol{x}_i = (x_{1i}, x_{2i}, ..., x_{Ki})$ the vector of observed baseline covariates for the i^{th} player. The full set of covariates is listed in Table 1. Besides the main players' achievements and some statistics of the relative team, two categorical variables were created: All-Star game (if the player was selected or not for playing at this competition) and G-League (if the player also played in the young championship).
Play-by-play data of the post-AS game season segment have been analyzed for extracting the needed information relative to the outcome variable. In detail, for each player, the minutes played until different time points (time referred to the second season segment - $s_1, ..., s_J$) and the corresponding scored points were collected. Let's denote with M_j and P_j, respectively, the variables relative to the minutes played until time s_j ($j = 1, 2, ..., J$) and the corresponding scored points. For each player, we recorded the amount of minutes m_{ij} played at time s_j, and the points p_{ij} gained after having played m_{ij} minutes (see Table 2). So, the time variable M_j was treated as player-time: m_{ij} increases when the player is in the court and remains constant when he is not playing.

Then, we fixed a given threshold P of scored points and we defined the event of interest as the exceeding of that threshold. Censoring occurred when the player did not exceed the fixed amount of points at the end of the post-AS game regular

Table 1: Baseline variables

Variable	Type
FGM - Field Goals Made	Numeric
FGA - Field Goals Attempted	Numeric
FG% - Percentage of Field Goals Made	Numeric
3PM - Three-Point Shots Made	Numeric
3PA - Three-Point Shots Attempted	Numeric
3P% - Percentage of Three-Point Shots Made	Numeric
2PM - Two-Point Shots Made	Numeric
2PA - Two-Point Shots Attempted	Numeric
2P% - Percentage of Two-Point Shots Made	Numeric
FTM - Free Throws Made	Numeric
FTA - Free Throws Attempted	Numeric
FT% - Percentage of Free Throws Made	Numeric
OREB - Offensive Rebounds	Numeric
DREB - Defensive Rebounds	Numeric
REB - Rebounds	Numeric
AST - Assists	Numeric
TOV - Turnovers	Numeric
STL - Steals	Numeric
BLK - Blocks	Numeric
PF - Personal Fouls	Numeric
FP - Fantasy Points	Numeric
DD2 - Double Doubles	Numeric
TD3 - Triple Doubles	Numeric
$+/-$ - Plus/Minus	Numeric
Age	Numeric
GP - Games Played	Numeric
Percentage Won matches (player)	Numeric
Percentage Loss matches (player)	Numeric
MIN - Minutes played	Numeric
Percentage won matches (team - per game)	Numeric
PTS (player) - Points gained (player)	Numeric
PTS (team) - Points gained (team - per game)	Numeric
All-Star game	Categorical (Yes-No)
NBA G-League	Categorical (Yes-No)

season segment. Two different settings with percentages of censoring equal to 20% and 50% were analyzed, corresponding to threshold values equal to 99 and 255 points, respectively. Moreover, an exploratory analysis has been performed for other thresholds and corresponding censoring percentages.

Finally, the response variable can be defined as follows. We denote with $j_i^*(P) = \text{argmin}_{j=1,\dots,J} p_{ij} > P$ the index j corresponding to the first time that the i^{th} player exceeds the threshold P. The time at which the player exceeds the threshold is therefore $s_{j_i^*(P)}$. Thus, the outcome of the study is composed of (i) the time-to-event $\tau_i = \min(m_{ij_i^*(P)}, m_{iJ})$, where $t_i = m_{ij_i^*(P)}$ is the amount of minutes played by the player for exceeding the threshold and $c_i = m_{iJ}$ corresponds to the amount of minutes played at the end of the season segment, and of (ii) the event indicator $\delta_i = I[p_{ij} > P]$. Then, the survival outcome for the i^{th} subject is

$$\tau_i = min(t_i, c_i) = \min(m_{ij_i^*(P)}, m_{iJ}) = \begin{cases} m_{ij_i^*(P)} & \text{if } \delta_i = I[p_{ij} > P] = 1 \\ m_{iJ} & \text{if } \delta_i = I[p_{ij} > P] = 0 \end{cases} . \quad (6)$$

Table 2: Example of collected data. Each row refers to a player, and each column to a time point. In each cell the overall amount of minutes played and points gained until that given time point are recorded

t_1	t_2	$\cdots$	t_J
(m_{11}, p_{11})	(m_{12}, p_{12})	$\cdots$	(m_{1J}, p_{1J})
$\vdots$	$\vdots$	$\ddots$	$\vdots$
(m_{n1}, p_{n1})	(m_{n2}, p_{n2})	$\cdots$	(m_{nJ}, p_{nJ})

4. RESULTS

4.1. PRELIMINARY ANALYSIS

The overall sample included 359 players, after having excluded those who changed team during the season and those who played less than 48 minutes in all the post-AS game season segment. Two distinct cases have been analyzed, with percentage of censoring 20% and 50% respectively, in order to examine if the covariates have a different impact on the outcome. The points' thresholds that allowed the desired censoring percentage were 99 and 255 for the setting with

20% and 50% censoring, respectively. Moreover, an exploratory analysis has been performed for other thresholds and corresponding censoring percentages.

The game variables denoting the total number of achievements of each player in the analyzed period have been normalized dividing by his minutes played (to have comparable results). Then, all the covariates have been standardized (to have mean 0 and standard deviation 1).

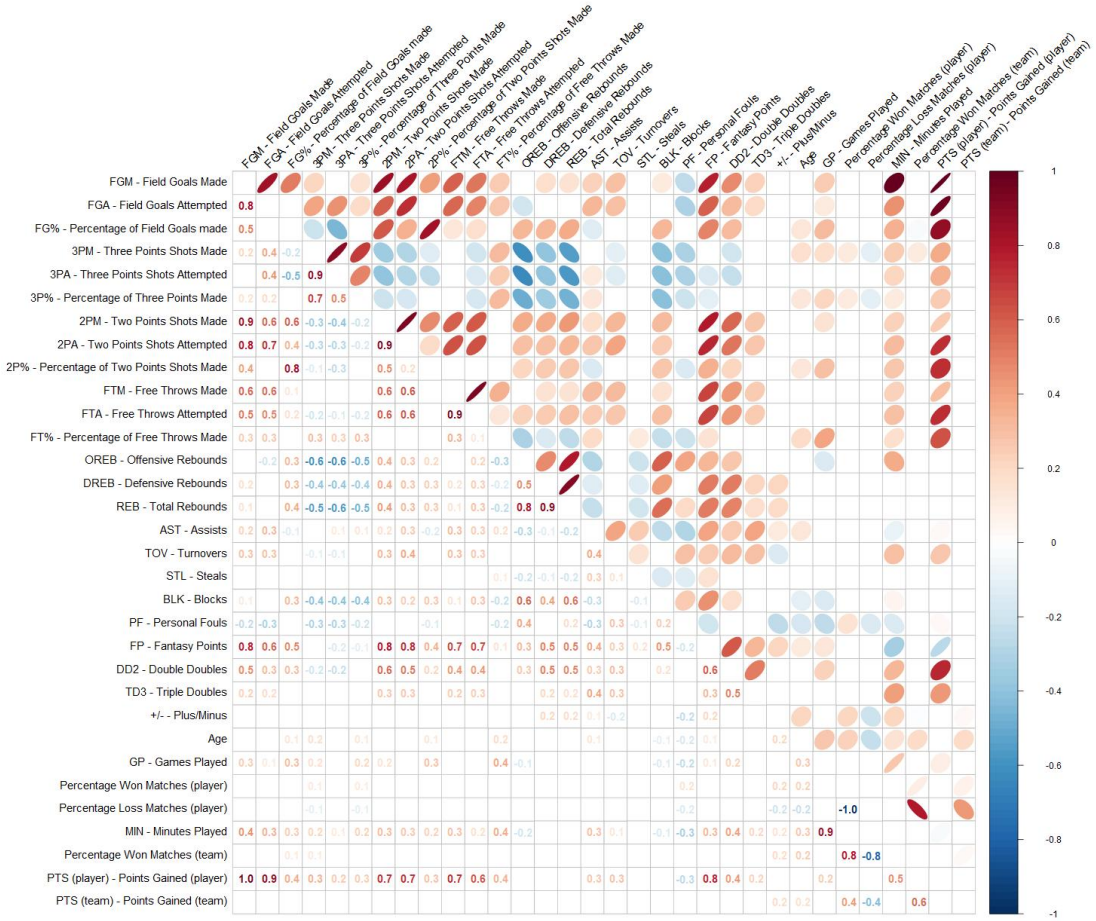
The first step of variable selection was carried out on the basis of prior knowledge and through the examination of correlations analysis. Indeed, considering all the NBA statistics would imply a high risk of multicollinearity, due to the presence of highly correlated variables (see Figure 1a). Therefore, we excluded some redundant variables (Figure 1b). After this step, the set of baseline covariates passed from 34 to 23. All these variables, as already pointed out, refer to the first season segment, i.e. the part of the regular season before the All-Star game. Among the excluded variables there is also the amount of fantasy points, due to the high multicollinearity found in another study (Macis et al., Submitted).

The following step of variable selection involved the use of Stepwise Cox regression model and its regularized version through the Lasso, for selecting the most important variables.

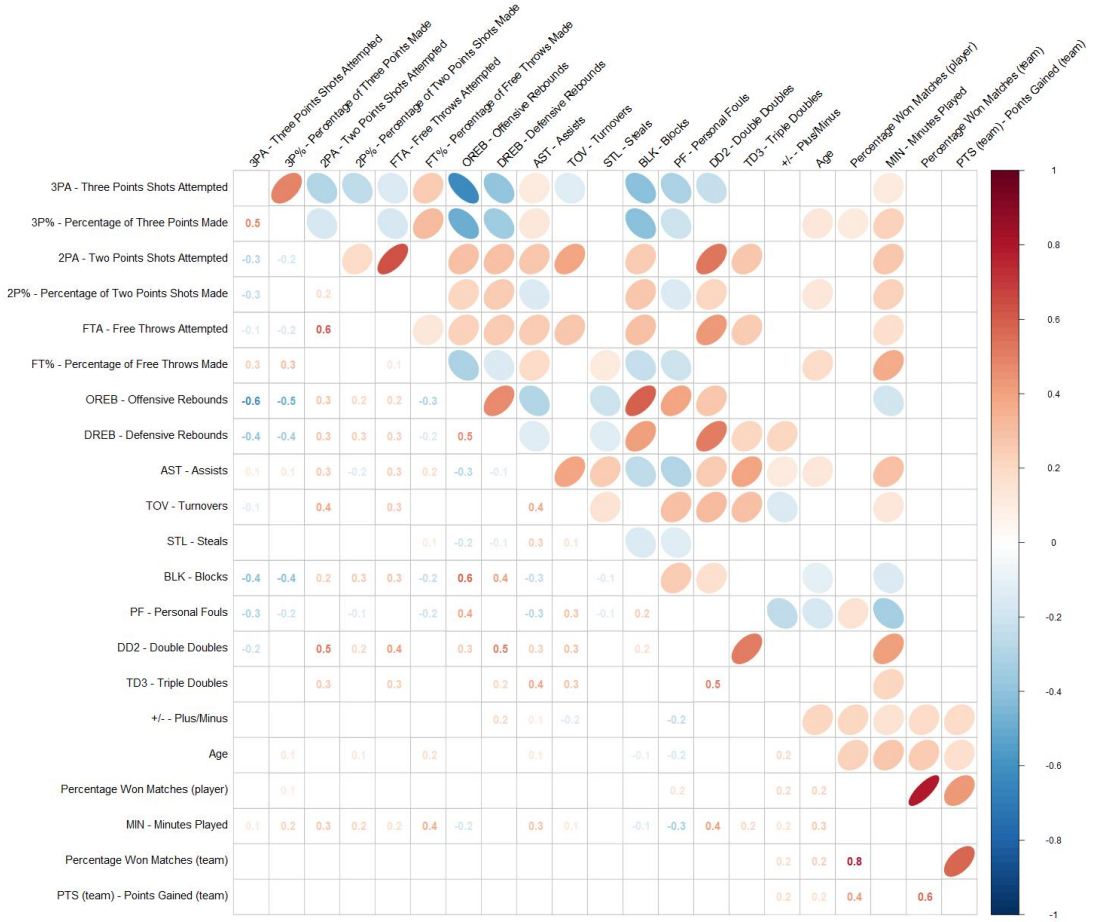
4.2. REGRESSION ANALYSIS

PERFORMANCE OF NBA PLAYERS

Using as a threshold 99 points, 287 players exceeded the given amount of points (corresponding to the 80% of the sample). Table 3 shows the results obtained fitting the two models. The Stepwise Cox model and the Lasso identified almost the same variables (with the exception of the All-Star game variable). More in detail, both models selected the amount of minutes played in the pre-AS season segment, the number of attempted shots (free throws -FTA-, two- -2PA- and three- -3PA- pointers), the percentage of two-point shots made (2P%) and the number of gained double doubles (DD2). All these variables resulted positively associated with the outcome: an increase in the number of these achievements is associated to a higher probability of exceeding the threshold. Moreover, the Stepwise Cox regression model and the Lasso Cox also identified the number of steals (STL), even if the estimated hazard ratio was not statistically significant in Stepwise Cox model ($p = 0.144$) and in the Lasso it seems to have a low estimated impact on the outcome (HR approximately equal to 1.00). Finally, the Lasso also identified the All-Star game variable, even if with an estimated HR approximately equal to one.



(a) Full set of covariates



(b) Set of covariates after the first step of variable selection

Figure 1: Correlation plot of the a) full set of covariates b) set of covariates obtained after the first step of variable selection. The game variables denoting the total number of achievements of each player have been normalized dividing by his minutes played; all the covariates have been standardized. Ellipses towards left (right) indicate negative (positive) correlations

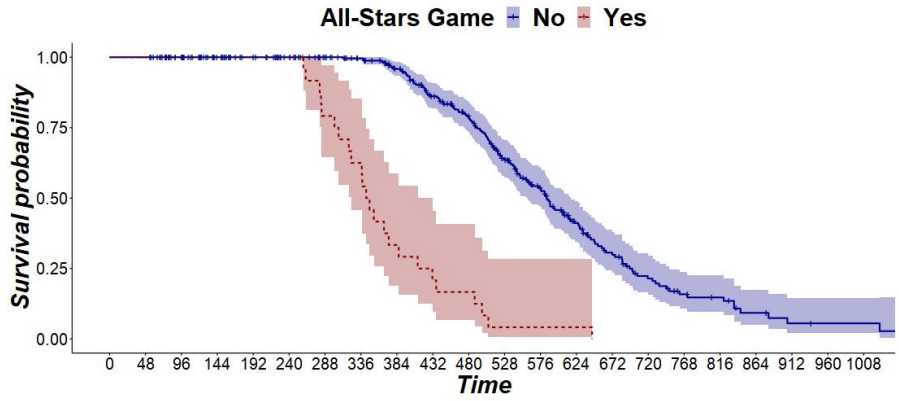
Increasing the threshold to 255 points, only one half (179 players) of the sample exceeded the points' cut-off. In this setting almost all the variables identified in the previous one were selected by the two models, but with some differences (Table 3). More in detail, the amount of minutes played in the previous season segment and the number of FTA were only identified by the Lasso with an estimated HR close to 1.00, suggesting a lower impact on the outcome. The 2P%, instead, was only identified by the Stepwise Cox model. Moreover, increasing the threshold of points, the number of STL was found negatively associated with the outcome. The estimated coefficients resulted lower than 1 (equal to 0.77 and 0.95 in Stepwise Cox and Lasso respectively), indicating that this achievement is negatively associated with the outcome: a unit increase of it leads to a lower probability of exceeding the threshold. Finally, with a higher threshold, the All-Star game resulted to be a relevant feature identified by both the models: having been selected for playing at the All-Star game doubles the probability of gaining more than 255 points.

Interestingly, most of the variables identified in this setting have a higher effect (higher HR) than in the setting with a lower threshold.

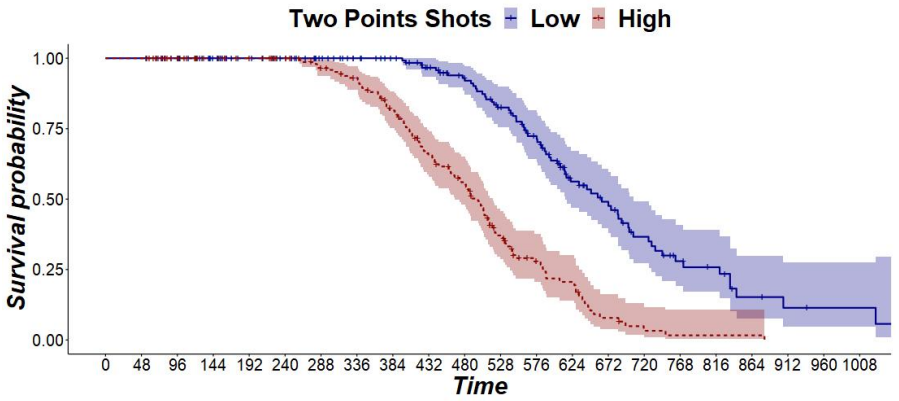
Figure 2 shows two examples of the estimated survival curves of the sample stratifying for two of the most important variables for the setting with a 50% of censoring. The number of 2PA (normalized and standardized) has been dichotomized in two categories with respect to the median. Figure 2a shows that players selected for the All-Star game reach the fixed amount of points very earlier than those who have not been selected for the match. Similarly, Figure 2b shows that players who attempted a higher number of two-point shots in the first part of the season have a higher probability of gaining the given threshold earlier than those who attempted a lower number of shots.

COEFFICIENTS AND PERCENTAGE OF CENSORING

Finally, an analysis of the pattern of the estimated coefficients as the percentage of censoring varies has been carried out. The results are shown in Figure 3. Each subfigure reports the estimated hazard ratios (i.e. $e^{\hat{\beta}}$) of each variable for different censoring settings (x-axis) for both the Stepwise Cox and the Lasso Cox. In details, we analyzed the censoring percentages ranging from 10% to 75% with a step of 5%. It can be seen that, almost always, the hazard ratios estimated with the Stepwise Cox are greater than those obtained by the Lasso Cox. The most important variables are the number of attempted two- and three-point shots (2PA and 3PA), the percentage of two-point shots (P2%), the number of double dou-



(a) All-Star game



(b) Two-Point Shots

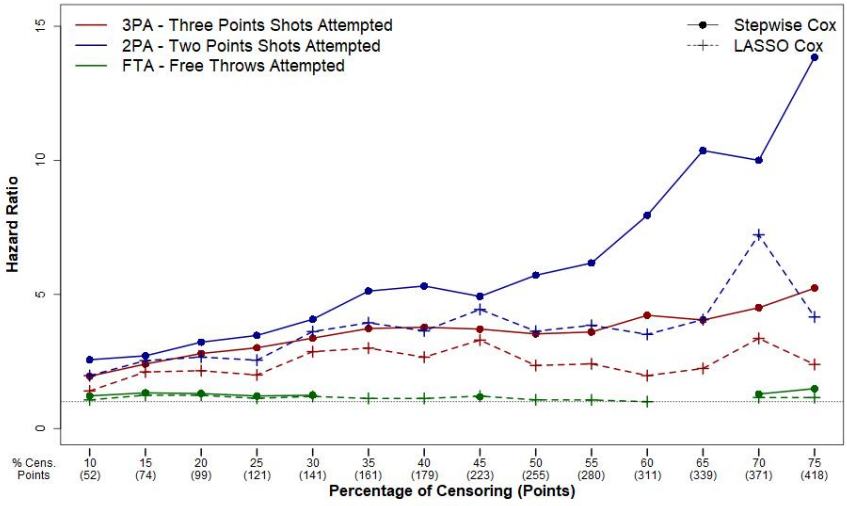
Figure 2: Survival Curves of the sample stratified for: a) All-Star game and b) Number of attempted two-point shots. The 2PA variable (normalized and standardized) has been dichotomized with respect to the corresponding median. The dashed line refers to the selection of the All-Star game and to the high category of 2PA, the solid line refers to having not been selected for the All-Star game and to the low category of 2PA.

Table 3: Results of the variable selection procedure in both the two settings (20% and 50% of censoring)

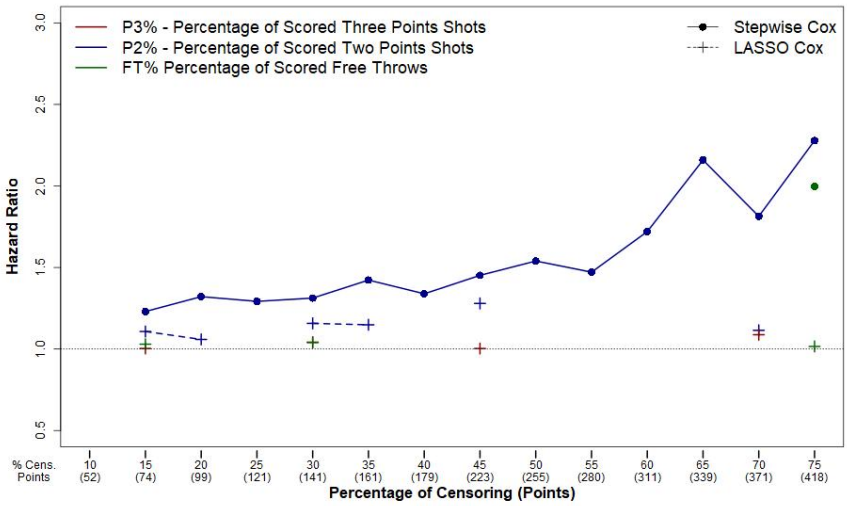
	20% censoring		50% censoring	
Variables Included in the Model	Stepwise HR (p)	Lasso HR	Stepwise HR (p)	Lasso HR
3PA - 3-Point Shots attempted	2.80 (< 0.001)	2.16	3.53 (< 0.001)	2.35
3P% - % 3-Point Shots made				
2PA - 2-Point Shots attempted	3.22 (< 0.001)	2.66	5.72 (< 0.001)	3.63
2P% - % 2-Point Shots made	1.32 (0.004)	1.06	1.54 (0.003)	
FTA - Free Throws attempted	1.30 (0.002)	1.24		1.07
FT% - % Free Throws made				
OREB - Offensive Rebounds				
DREB - Defensive Rebounds				
AST - Assists				
TOV - Turnovers				
STL - Steals	0.90 (0.144)	1.00	0.77 (0.018)	0.95
BLK - Blocks				
PF - Personal Fouls				
DD2 - Double Doubles	1.27 (0.001)	1.21	1.27 (0.003)	1.21
TD3 - Triple Doubles				
$+/-$ - Plus/Minus				
Age				
% Won Matches (by the player)				
MIN - Minutes played	1.24 (0.004)	1.20		1.08
% Won Matches (by the team)				
Points Gained (by the team)				
All-Star Game (Yes/No)		1.05	2.27 (0.004)	1.86
NBA G-League (Yes/No)				

bles (DD2) and the number of steals (STL). These variables are selected in almost all the settings by the two models. Moreover, the All-Star game and the number of minutes played are selected many times by the Stepwise Cox and always by Lasso Cox. As the percentage of censoring increases, the estimated hazard ratios increase. On the other hand, the estimated hazard ratio for STL is always negative and its value decreases as the fixed threshold (and consequently the percentage of

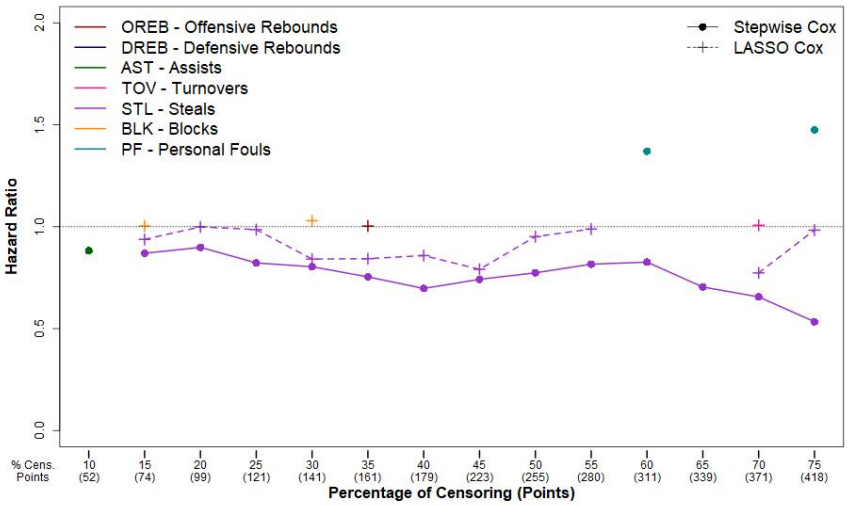
censoring) increases.



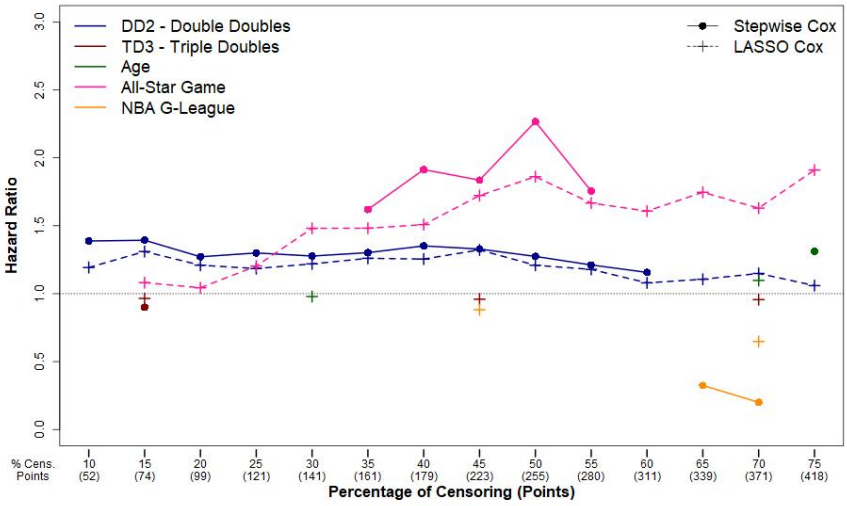
(a) Attempted Shots



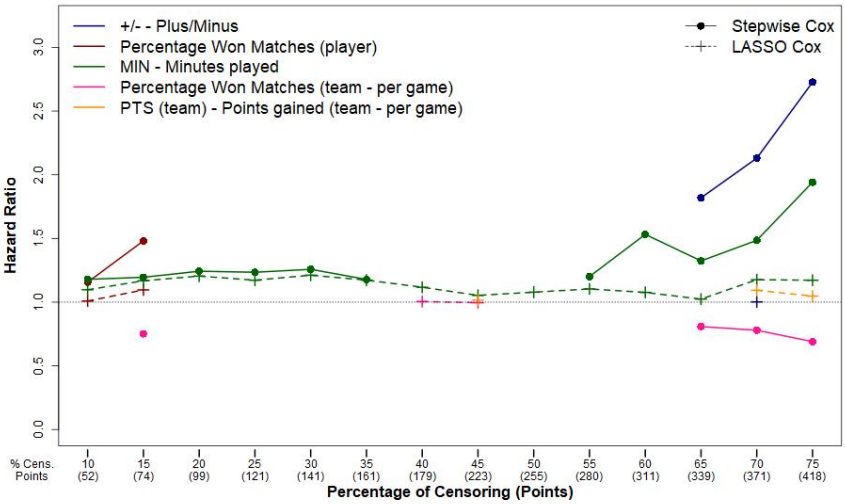
(b) Percentage of realized shots



(c) Defense and Game construction



(d) Achievements



(e) Points and Winning Percentage

Figure 3: Estimated hazard ratios for different censoring percentages

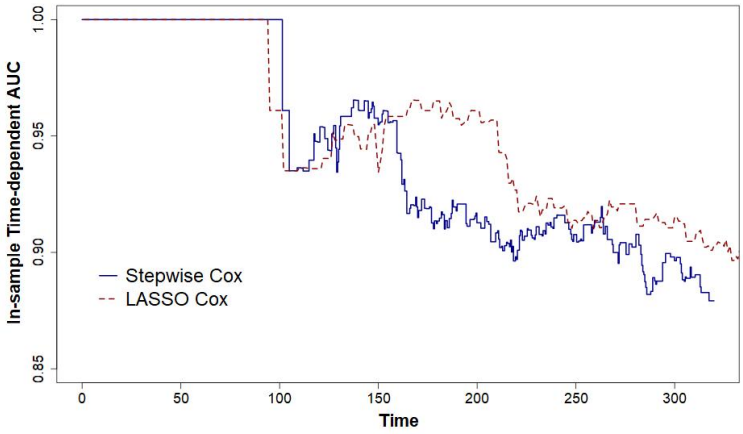
4.3. EVALUATION OF MODELS' PERFORMANCE

In order to evaluate the performance of the Lasso Cox model and test the assumption of proportionality of hazards, a Cox model with the variables selected by Lasso was fitted.

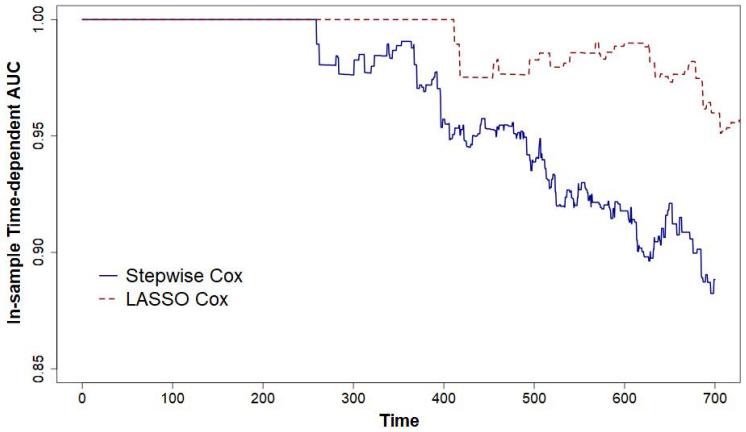
The two models satisfied the assumption of proportionality of hazards (null hypothesis) in both settings, as measured by the statistical test based on Shoenfield residuals ($p = 0.072$ and 0.077, respectively, for Stepwise and Lasso Cox models, when the percentage of censoring was equal to 20%, and $p = 0.093$ and 0.077 when the percentage of censoring was equal to 50%).

Then, model's performance has been evaluated through the evaluation of time-dependent AUCs. The time-dependent AUC measures the area under the ROC curve evaluated at different timepoints; thus, it differs from classical ROC analysis because the outcome of an observation can change over time and because of the presence of censoring. The time-dependent AUC assesses the ability of the model to discriminate the binary outcome (event/non event) at different timepoints. Values close to one indicate a good performance. The performance assessment has been made in-sample, so the performance could have been overestimated. It can be seen that in both the settings the performance of the two models

is good (greater than 0.85). Moreover, it can be observed that the Lasso seems to slightly overperform Stepwise Cox in both the settings (Figure 4).



(a) 20% **censoring**



(b) 50% **censoring**

Figure 4: Time-dependent AUC for Stepwise Cox and Lasso Cox in the two settings. a) 20% **of censoring. b)** 50% **of censoring.**

5. DISCUSSION

Survival analysis has been already used for sport analytics for many aims; however, up to now, to the best of our knowledge, it has never been used for evaluating players' performance. This study shows the use of a classical method of survival analysis, as Stepwise Cox regression, and of a more recent extension, regularized Cox regression through Lasso, for identifying the achievements that are highly associated to the offensive performance of NBA players, measured in terms of exceeding of a given threshold of points. Two settings were analyzed, with thresholds equal to 99 and 255 points, corresponding to censoring percentages of 20% and 50%, respectively, for examining whether there is a different impact of the considered variables on the outcome. Other reasonable values for the percentage of censoring that can be investigated are those ranging from 10% to about 75%, as shown in Figure 3. Values higher than 75% could instead lead to possible non-meaningful results, because in these cases it is more likely that the follow-up is too short with respect to the event under analysis. However, besides the meaning of the results obtained from the analyses, attention has to be paid to the proportional hazard assumption, which may not be respected in all the settings.

Summarizing, in the two examined settings almost all the same variables were selected by both the models. In particular, from both Stepwise Cox regression and Lasso, it emerged that (as expected) players who attempted more shots, free throws, $2-$ and $3-$point shots in the pre-AS game season segment have a higher probability of exceeding the fixed amount of points in the second part of the season. In particular, it emerged that the most important variable, i.e. the one with the highest estimated hazard ratios, is the number of 2PA, followed by the number of 3PA, and that their impact increases when the threshold is higher. Moreover, both the two models suggest that gaining more achievements (as measured by the number of DD2) is associated to an increase of the probability of success in a shorter time. In addition, the number of STL was identified as negatively associated to the outcome. Thus, it seems that this variable decreases the probability of reaching the two thresholds (HR < 1). Its importance is relevant when the threshold is high, while it is not statistically significant ($p > 0.05$) when the fixed threshold of points is low. This is an attractive result because it is the only variable related to defense included in the models. This may suggest that who is more involved in defense is then penalized in terms of scored points. On the other hand, variables like rebounds and blocks have not been selected by the models and, from this point of view, defense seems to be not influential on the scored points. These

remarks about the role of defense should be deepened with some research specifically devoted to answer this question.

Finally, an interesting finding emerged from the comparison of the results of the two settings; indeed, All-Star game becomes an important factor when considering a higher threshold. Therefore, when the fixed amount of points is higher, being a very good player (and so having been selected for playing at the All-Star game) almost doubles the probability of reaching that threshold (HR=2.3 and 1.9 for Stepwise Cox and Lasso respectively).

All these results are also confirmed by those shown in Figure 3 for different percentages of censoring.

An interesting idea is to also consider the features of the opponent teams, in order to verify whether the opponent teams have an impact on the players' performance. This could be done, for example, by weighting some of the covariates (e.g. 2PA and 2PM) with respect to the ranking of the corresponding opponent team, in order to also consider the possible impact of the team against which the achievements have been gained.

The study has some limitations associated to possible issues related to the assumption of *random censoring*. Random censoring occurs when the subjects who are censored at time t are representative of all the study subjects who remained at risk at time t, with respect to their survival experience (Kleinbaum et al., 2012). This hypothesis may be not respected due to the fact that it is likely that censored players (i.e. players who didn't exceed the threshold) have not the same abilities of players who manage to exceed that amount of points. On the other side, the assumption of *independent censoring* can be retained valid. Indeed, it is reasonable to assume that within any subgroup of interest, the subjects who are censored at time t are representative of all the subjects in that subgroup who remained at risk at time t with respect to their survival experience. So, random censoring could be assumed conditional on each level of covariates (Kleinbaum et al., 2012). For this reason, once having taken into account the abilities of each player, as measured through the covariates introduced in the model, the probability of being censored can be considered independent of the probability that the event of interest occurs. Future research will deepen this issue.

Moreover, due to the relatively low number of subjects, the full original sample has been used for fitting the model and performance has been evaluated in-sample. Future work will consider the use of data relative to 2021-2022 season as test set.

Finally, some improvements include considering the possible presence of in-

teractions among covariates, non-linear effects of the predictors and threshold effects. Non-parametric and machine learning methods will be used to examine this point.

ACKNOWLEDGMENTS

The authors thank the reviewers for their valuable comments, which greatly improved the paper, the Editor and the Guest Editors. Thanks go to BigDataball (`www.bigdataball.com`) for supplying us high quality sports data for scientific research.

This study was carried out in collaboration with the Big&Open Data Innovation Laboratory at the University of Brescia (project 'BDsports: Big Data analytics in sports'; `bdsports.unibs.it`).

References

Back, F.A., Hino, A.A.F., Bojarski, W.G., Aurélio, J.M.G., de Castro Moreno, C.R., and Louzada, F.M. (2022). Evening chronotype predicts dropout of physical exercise: a prospective analysis. In *Sport Sciences for Health*, 1–11.

Beynnon, B.D., Vacek, P.M., Murphy, D., Alosa, D., and Paller, D. (2005). First-time inversion ankle ligament trauma: the effects of sex, level of competition, and sport on the incidence of injury. In *The American journal of sports medicine*, 33 (10): 1485–1491.

Buist, I., Bredeweg, S.W., Bessem, B., Van Mechelen, W., Lemmink, K.A., and Diercks, R.L. (2010). Incidence and risk factors of running-related injuries during preparation for a 4-mile recreational running event. In *British journal of sports medicine*, 44 (8): 598–604.

Clarke, P.M., Walter, S.J., Hayen, A., Mallon, W.J., Heijmans, J., and Studdert, D.M. (2012). Survival of the fittest: retrospective cohort study of the longevity of Olympic medallists in the modern era. In *Bmj*, 345.

Collett, D. (2015). *Modelling survival data in medical research*. CRC press.

Cox, D.R. (1972). Regression Models and Life-Tables. In *Journal of the Royal Statistical Society. Series B (Methodological)*, 34 (2): 187–220. URL `http://www.jstor.org/stable/2985181`.

Csurilla, G. and Fertő, I. (2022). How long does a medal win last? Survival analysis of the duration of Olympic success. In *Applied Economics*, 1–15.

Dekker, T.J., Godin, J.A., Dale, K.M., Garrett, W.E., Taylor, D.C., and Riboh, J.C. (2017). Return to sport after pediatric anterior cruciate ligament reconstruction and its effect on subsequent anterior cruciate ligament injury. In *JBJS*, 99 (11): 897–904.

Del Corral, J., Barros, C.P., and Prieto-Rodriguez, J. (2008). The determinants of soccer player substitutions: A survival analysis of the Spanish soccer league. In *Journal of Sports Economics*, 9 (2): 160–172.

Derksen, S. and Keselman, H.J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. In *British Journal of Mathematical and Statistical Psychology*, 45 (2): 265–282.

Ekeland, A., Engebretsen, L., Fenstad, A.M., and Heir, S. (2020). Similar risk of ACL graft revision for alpine skiers, football and handball players: the graft revision rate is influenced by age and graft choice. In *British Journal of Sports Medicine*, 54 (1): 33–37.

Ekman, A. (2017). Variable selection for the Cox proportional hazards model: A simulation study comparing the stepwise, lasso and bootstrap approach.

Fynn, K.D. and Sonnenschein, M. (2012). An analysis of the career length of professional basketball players. In *The Macalester Review*, 2 (2): 3.

Gutiérrez, E., Lozano, S., and González, J.R. (2011). A recurrent-events survival analysis of the duration of Olympic records. In *IMA Journal of Management Mathematics*, 22 (2): 115–128.

Harrell, F.E. (2015). Regression modeling strategies with applications to linear models, logistic and ordinal regression, and survival analysis.

Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., and Rosati, R.A. (1982). Evaluating the yield of medical tests. In *Jama*, 247 (18): 2543–2546.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). Statistical learning with sparsity. In *Monographs on statistics and applied probability*, 143: 143.

Hopkins, W.G., Marshall, S.W., Quarrie, K.L., and Hume, P.A. (2007). Risk factors and risk statistics for sports injuries. In *Clinical Journal of Sport Medicine*, 17 (3): 208–210.

Howell, D.R., Potter, M.N., Kirkwood, M.W., Wilson, P.E., Provance, A.J., and Wilson, J.C. (2019). Clinical predictors of symptom resolution for children and adolescents with sport-related concussion. In *Journal of Neurosurgery: Pediatrics*, 24 (1): 54–61.

Jack, R.A., Sochacki, K.R., Hirase, T., Vickery, J., McCulloch, P.C., Lintner, D.M., and Harris, J.D. (2019). Performance and return to sport after hip arthroscopic surgery in major league baseball players. In *Orthopaedic Journal of Sports Medicine*, 7 (2): 2325967119825835.

Kleinbaum, D.G., Klein, M., et al. (2012). *Survival analysis: a self-learning text*, vol. 3. Springer.

Kontos, A.P., Elbin, R., Sufrinko, A., Marchetti, G., Holland, C.L., and Collins, M.W. (2019). Recovery following sport-related concussion: integrating pre-and postinjury factors into multidisciplinary care. In *The Journal of Head Trauma Rehabilitation*, 34 (6): 394–401.

Lawrence, D.W., Richards, D., Comper, P., and Hutchison, M.G. (2018). Earlier time to aerobic exercise is associated with faster recovery following acute sport concussion. In *PLoS One*, 13 (4): e0196062.

Lu, Y., Jurgensmeier, K., Till, S.E., Reinholz, A., Saris, D.B., Camp, C.L., and Krych, A.J. (2022). Early ACLR and Risk and Timing of Secondary Meniscal Injury Compared With Delayed ACLR or Nonoperative Treatment: A Time-to-Event Analysis Using Machine Learning. In *The American Journal of Sports Medicine*, 03635465221124258.

Macis, A., Manisera, M., Sandri, M., and Zuccolotto, P. (Submitted). Which Achievements Are Associated To a Better Offensive Performance in NBA? A Survival Analysis Study. In *13th World Congress of Performance Analysis of Sport (WCPAS2022) the 13th International Symposium on Computer Science in Sport (IACSS2022), Proceedings*, —. Springer.

Mahmood, A., Ullah, S., and Finch, C. (2014). Application of survival models in sports injury prevention research: a systematic review. In *British journal of sports medicine*, 48 (7): 630–630.

Mai, H.T., Chun, D.S., Schneider, A.D., Erickson, B.J., Freshman, R.D., Kester, B., Verma, N.N., and Hsu, W.K. (2017). Performance-based outcomes after

anterior cruciate ligament reconstruction in professional athletes differ between sports. In *The American journal of sports medicine*, 45 (10): 2226–2232.

Moulds, K., Abbott, S., Pion, J., Brophy-Williams, C., Heathcote, M., and Cobley, S. (2020). Sink or swim? A survival analysis of sport dropout in Australian youth swimmers. In *Scandinavian journal of medicine & science in sports*, 30 (11): 2222–2233.

Nelson, L.D., Tarima, S., LaRoche, A.A., Hammeke, T.A., Barr, W.B., Guskiewicz, K., Randolph, C., and McCrea, M.A. (2016). Preinjury somatization symptoms contribute to clinical recovery after sport-related concussion. In *Neurology*, 86 (20): 1856–1863.

Nevo, D. and Ritov, Y. (2013). Around the goal: examining the effect of the first goal on the second goal in soccer using survival analysis methods. In *Journal of Quantitative Analysis in Sports*, 9 (2): 165–177.

Pion, J., Lenoir, M., Vandorpe, B., and Segers, V. (2015). Talent in female gymnastics: a survival analysis based upon performance characteristics. In *International journal of sports medicine*, 94 (11): 935–940.

Pratas, J.M., Volossovitch, A., and Carita, A.I. (2016). The effect of performance indicators on the time the first goal is scored in football matches. In *International Journal of Performance Analysis in Sport*, 16 (1): 347–354.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Coxâ s proportional hazards model via coordinate descent. In *Journal of statistical software*, 39 (5): 1.

Smith, K.L. and Weir, P.L. (2022). An Examination of Relative Age and Athlete Dropout in Female Developmental Soccer. In *Sports*, 10 (5): 79.

Sochacki, K.R., Jack, R.A., Hirase, T., Vickery, J., and Harris, J.D. (2019). Performance and return to sport after hip arthroscopy for femoracetabular impingement syndrome in National Hockey League players. In *Journal of Hip Preservation Surgery*, 6 (3): 234–240.

Tay, K., Simon, N., Friedman, J., Hastie, T., Tibshirani, R., and Narasimhan, B. (2022). Regularized Cox Regression.

Thomas, A.C. (2007). Inter-arrival times of goals in ice hockey. In *Journal of Quantitative Analysis in Sports*, 3 (3).

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. In *Statistics in medicine*, 16 (4): 385–395.

Tozetto, A.B., Carvalho, H.M., Rosa, R.S., Mendes, F.G., Silva, W.R., Nascimento, J.V., and Milistetd, M. (2019). Coach turnover in top professional Brazilian football championship: A multilevel survival analysis. In *Frontiers in psychology*, 10: 1246.

Venturelli, M., Schena, F., Zanolla, L., and Bishop, D. (2011). Injury risk factors in young soccer players detected by a multivariate survival model. In *Journal of science and medicine in sport*, 14 (4): 293–298.

Wangrow, D.B., Schepker, D.J., and Barker III, V.L. (2018). Power, performance, and expectations in the dismissal of NBA coaches: A survival analysis study. In *Sport Management Review*, 21 (4): 333–346.

Zuccolotto, P. and Manisera, M. (2020). *Basketball data science: with applications in R*. CRC Press.

Zumeta-Olaskoaga, L., Weigert, M., Larruskain, J., Bikandi, E., Setuain, I., Lekue, J., Küchenhoff, H., and Lee, D.J. (2021). Prediction of sports injuries in football: a recurrent time-to-event approach using regularized Cox models. In *AStA Advances in Statistical Analysis*, 1–26.