Predicting full retirement attainment of NBA players

Foutzopoulos, Giorgos and Pandis, Nikolaos and Tsagris, Michail

Independent Researcher, University of Bern, University of Crete

23 July 2024

Predicting full retirement attainment of NBA players

Foutzopoulos Giorgos[1], Pandis Nikolaos[2] and Tsagris Michail[3]

[1] Independent researcher,Athens,Greece
gfoutzopoulos@gmail.com
[2] Department of Orthodontics and Dentofacial Orthopedics,
Dental School/Medical Faculty, University of Bern, Bern, Switzerland
npandis@yahoo.com
[3] Department of Economics, University of Crete, Gallos Campus, Rethymnon, Greece
mtsagris@uoc.gr

July 23, 2024

Abstract

The aim of this analysis is to predict whether an National Basketball Association (NBA) player will be active in the league for at least 10 years so as to be qualified for NBA's full retirement scheme which allows for the maximum benefit payable by law. We collected per game statistics for players during their second year, drafted during the years 1999 up to 2006, for which, information on their career longetivity is known. By feeding these statistics of the sophomore players into statistical and machine learning algorithms we select the important statistics and manage to accomplish a satisfactory predictability performance. Further, we visualize the effect of each of the selected statistics on the estimated probability of staying in the league for more than 10 years. Finally, as an illustration, we collected data from players that were drafted 11 years ago (and some are still active) and estimated their probability of surviving in the league for at least 10 years.

Keywords: NBA, career duration, exit discrimination, retirement scheme
JEL Codes: C21, C38, C41, C53

1 Introduction

Predicting career longevity from the early stages as professional athletes is an interesting task, both from the athlete's and the club's perspective as it is in the interest of both sides to decide on future strategy. The task becomes challenging, and more interesting at the same time, when one relies on early stage professional performance statistics. Although there are several research papers on this topic, in this paper, we will narrow our attention to U.S. professional basketball athletes competing in the NBA league. Specifically, we will focus on whether an NBA player survives in the league for at least 10 years, which is the minimum number of years in order to receive the full retirement scheme.

Staw and Hoang (1995) delved into factors influencing NBA players' longevity, analyzing data from the 1980-1986 drafts until the 1990-1991 season. Their event history analysis identified a player's initial draft number, performance variables, and tenure length as pivotal factors affecting career longevity, with scoring ability significantly impacting playing time and the likelihood of remaining in the league. The study also revealed that NBA franchises tended to retain high draft choices over low draft choices, while defensive skills like rebounding and blocked shots positively influenced a player's chances of staying in the league, particularly within teams valuing team-oriented skills. In parallel, Groothuis and Hill (2004) and Groothuis and Hill (2018) conducted significant research on exit discrimination in the NBA, focusing on non-U.S. players. Groothuis and Hill (2004) explored factors affecting career duration, emphasizing team owners' inclination to retain productive players, with assists, blocks, and points per minute played influencing continued NBA tenure. Height, weight, and draft number were also identified as significant predictors of career length, while race did not contribute to exit discrimination. Groothuis and Hill (2018) expanded on this, by revealing that foreign-born players without U.S. college experience tended to have shorter careers, possibly indicating exit discrimination or a preference for concluding their careers in their home countries. Conversely, players with U.S. college experience exhibited career lengths comparable to native-born players, highlighting the complex interplay of fan preferences, cultural dynamics, and lucrative opportunities in shaping the career trajectories of foreign-born NBA players. These insights contribute significantly to sports economics and the broader discourse on international talent dynamics in professional sports leagues.

Petersen et al. (2011) demonstrated the Matthew effect ("rich get richer") where an individual's longevity and past success contribute to further career advancement. The study effectively illustrated that even a modest rate of progress at the onset of one's career has a crucial role in shaping the trajectory of career length. The model intricately incorporated the Matthew effect, underscoring the critical significance of early career development. This work shed light on the inherent disparities between short and protracted careers, revealing a compelling statistic that approximately 3% of basketball players experience an NBA career when playing for less than 12 minutes per game. Furthermore, the research accentuated that athletes enjoying extended careers successfully sustained a high level of performance over a substantial interval of playing time.

Two studies published in 2008 explored the significance of college basketball in shaping the trajectory of an NBA player's career. Coates and Oguntimein (2010) focused on NBA draft classes from 1987 to 1989 evaluated the predictability of successful careers based on college performance by examining retired players. The analysis, incorporating comprehensive data on draft details and performance metrics, revealed that players from smaller conferences exhibited higher efficiencies, driven by superior college points and rebounds. Despite achieving similar NBA production, players from smaller conferences experienced shorter careers compared to their counterparts from larger conferences, challenging prevailing notions about statistical discrimination and option value. By exploring correlations between college and NBA performance, Coates and Oguntimein (2010) provided valuable insights into the intricacies of draft decisions and player career trajectories. Barnes (2008) investigated the relationship between pre-NBA career statistical variables and NBA player longevity, conditioning on the players' playing positions, guard, forward, and center. Analyzing data from the 1988–2002 collegiate seasons, they employed 11 independent variables such as points, assists, and turnovers, with career longevity being the dependent variable. The statistical analysis unveiled significant associations for guards and forwards, emphasizing the impact of assists, turnovers, points, field goal percentage, and free throw percentage on NBA career longevity. Notably, the study found statistical insignificance for centers, attributing it to the unique nature of the center position and a smaller sample size. These findings underscore the potential of statistical analysis in assisting NBA general managers and scouts in effective player evaluation and selection strategies, contributing valuable insights into the complex process of building successful basketball teams. Miguel et al. (2019) performed an extensive analysis of NBA draft data from 1978 to 1998, revealing compelling insights into the relationship between draft selection order and players' career longevity. Players chosen in the first five picks, on average, enjoyed a more extended career of around 14 years, with a discernible non-linear trend showing a decrease in longevity from the first to the 30th pick. When accounting for draft years, the study identified fluctuations in career longevity, with an increase until 1985, stabilization until 1993, and a subsequent rise.

Fynn and Sonnenschein (2012) departed from conventional player performance metrics, opting instead for individual awards as a measure of success, they employed the number of individual awards won as a measure of performance, along with the player's biological data such as height and weight. They pointed out that, a player's height and number of awards won have a positive effect on his career duration. The association between height and extended career duration can be attributed to the scarcity of players in positions like Center and Forward-Center compared to guards or guard-forwards, making the former more sought after for their abilities in finishing shots around the rim, rebounding, and shot-blocking, regardless of specific performance metrics.

Career longevity is further contingent upon various factors, with season injuries and illnesses playing a significant role in determining the career span of an NBA player, as evidenced in the following studies. Kester et al. (2017) conducted a thorough investigation into the impact of anterior cruciate ligament (ACL) injury tears on NBA players from 1984 to 2014. Despite an 86.1% return rate post-ACL reconstruction, the study revealed a significantly shorter mean post-operative play of 1.84 years compared to controls. Survival analysis emphasized a heightened rate of early attrition for players undergoing ACL reconstruction, highlighting the intricate relationship between these injuries, rehabilitation success, and the enduring consequences on professional basketball players' career longevity. Khalil et al. (2020) using matched controls examined the consequences of Achilles tendon (AT) ruptures on NBA players' careers from 1970 to 2019 showed that among the 47 players with AT ruptures, an impressive 72.3% successfully resumed NBA participation post-surgery, albeit with significantly shortened playing careers compared to uninjured counterparts (3.1 vs. 5.8 seasons on average, respectively). Johns et al. (2021) conducted a review that examined the impact of Achilles tendon (AT) rupture on 333 professional athletes across major sports leagues. Findings reveal a 76.4% return-to-play rate after AT repair, with an average recovery time of 11 months—twice that of the general population. However, returning athletes experienced a significant decline in performance, particularly in the NFL and NBA, suggesting a potential career-altering consequence. That study underscored these athletes' challenges, providing crucial insights for setting evidence-based expectations in postoperative return to professional sports.

Martin et al. (2021) focused specifically on injuries during the rookie season of an NBA player. Using data from 2007 to 2019, they revealed heightened injury and illness rates in rookie players, particularly in the ankle. They explored the connection between rookie season injuries and career longevity. The results showed a significant reduction in total seasons played for rookies with injuries, but this effect lessened after accounting for confounding variables. Lower draft positions associated with shorter NBA careers, suggesting performance factors and organizational investments play a role. Specific injury patterns, notably ankle and knee injuries, emphasized the long-term consequences and advocated for targeted mitigation programs. While rookies exhibited a higher injury risk, adjusted analyses indicated career longevity is multi-factorial, with cumulative injury burden emerging as a potential determinant, emphasizing the need for ongoing research and improved mitigation strategies.

The goal of this paper deviates from the previous research works in that instead of attempting to predict the duration of NBA players it attempts to predict the likelihood of staying in the league for at least 10 years. Players who have served in the league for at least 3 years are eligible for the NBA's minimum pension package, but those who have served for 10 years are entitled to a full pension scheme that includes all possible benefits. NBA players can start receiving smaller monthly payments, over an extended period of time, as early as 45 years of age under the NBA Early Retirement Day scheme. Players are encouraged to hold off on receiving payments until the Normal Retirement Day at the age of 62 year to receive the highest possible payments. For instance, a player with only three years of service who opts into the pension at age 62 will receive the minimum amount of 56,998 dollars per year and a player with at least 10 years of service can get up to 215,000 dollars annually at the age of 62 years. The pension amount is based on a combination of factors that include years of service, age, and salary history hoopshype.com. It is worth highlighting that from July 2023 and on the monthly amount per Year of Credited Service payable as a Normal Retirement Pension is $1,001.47.

Statistical and machine learning (ML) algorithms were employed to this end, using data collected from drafted players between 1999 and 2006. Their second year performance statistics were fed into the algorithms achieving a high predictive performance. The same analysis was conducted again, but this time by retaining the most important statistics, and the predictive performance remained at the same levels of accuracy. As an extra validation scheme of the optimal model selected, more recent data were collected, exhibiting a very satisfactory performance of the model. The model is simple to use and interpret and could act as a first prediction of a player's career longetivity, while showing the effect of each performance on the probability of suvriving in the league for at least 10 years.

1.1 Pension benefits

It is noteworthy to highlight the significant shift in player eligibility for pension benefits between the old Collective Bargaining Agreement (CBA) of 2018-2023 and the updated agreement that came into effect in July 2023[1]. Under the previous CBA, a player was considered to have completed a full season simply by participating or being active in just one game. This rule applied even to two-way players, whose salaries for NBA workdays were included in the Total Salaries and Benefits, consequently contributing to the players' share of Basketball Related Income (BRI). However, with the implementation of the new CBA, effective July 2023, the criteria for a player to be considered on a roster underwent a significant revision. Now, a player is deemed to be on a roster if they are listed as active, inactive, or on a two-way list of any team on February 2nd of the ongoing regular season, or if they are on the active list for at least fifty percent (50%) of the total regular season games played by the team. This shift in eligibility criteria particularly impacts players in the second year of their professional career, as under the old CBA rule until the 2022-2023 season, they could become eligible for a pension by merely being under contract for at least one game in their third year in the league. Similarly, two-way contract players could qualify for pension benefits by being on an NBA team's roster for just one game.

Furthermore, as part of the 2017 CBA agreement, significant enhancements were made to the healthcare and educational benefits available to NBA players. Notably, retirees with a minimum of three years of service in the NBA now receive lifelong healthcare coverage, a provision unparalleled in other retiree associations. Moreover, those who have served for a decade or more in the league enjoy comprehensive healthcare coverage not only for themselves but also for their spouses and children. This comprehensive healthcare program sets a new standard in professional sports, ensuring that retired NBA players and their families are well taken care of for life. In addition to healthcare benefits, the CBA also introduced provisions for educational support. Retired players who wish to pursue further education can have their tuition reimbursed, up to $33,000 annually, changed to $62,500 for each calendar year on the 2023 CBA. This assistance aims to facilitate the transition to post-basketball careers and encourage lifelong learning. The educational benefits extend even further, as each eligible player with three (3) or more Years of NBA Service as of the date of the 2023 CBA Agreement shall receive a one-time increase in Tuition Reimbursement Benefit equal to $24,000 to either complete the degree if unfinished or pursue further studies. As of the latest statistics available as of September 2019, 28 players have already been approved for tuition reimbursement, with over 50 more awaiting approval. These initiatives underscore the NBA's commitment to supporting its players beyond their playing careers, promoting their overall well-being and continued personal development NBA-NBPA Collective Bargaining Agreement 2017.

It is also important to note that the NBA introduced a robust 401(k) benefits plan in the 2011 CBA[2], which was subsequently restated in both the 2017 and 2023 CBAs, tailored specifically for its players. This initiative provides a structured pathway for financial security during and after their playing careers. Under this plan, players can allocate a portion of their earnings into the 401(k) account, with contributions made on a pre-tax basis, effectively reducing their taxable income and allowing invetsments to grow tax free. In addition, the NBA as an employer matches these contributions, offering up to 140 percent of the player's own contributions. This generous matching scheme serves as a compelling encouragement for players to prioritize long-term financial planning. Within the 401(k) plan, players are presented with a diverse array of investment options, empowering them to tailor their investment strategies to their individual financial goals and risk tolerance. Furthermore, the structure of the NBA's 401(k) plan ensures disciplined saving habits, as players generally cannot access funds until they reach the age of 59-60 without incurring penalties. This safeguard is designed to fortify players

[1] NBA-NBPA CBA 2023
[2] NBA-NBPA CBA 2011

walking away from the game with significantly more money in savings hoopshype.com.

A description of the data collected, an some exploratory data analysis are presented in the next section, while Section 3 describes the analysis, the identification of the most important performance statistics and contains the results. Section 4 illustrates the performance of the final model to more recent data and finnaly Section 5 concludes the paper.

2 Description of the data

The sample consists of players who were drafted while attending a U.S. college or upon completion an undergraduate program. Players who were drafted outside a U.S. college, e.g. professional players from Europe or other continents were excluded from the analysis. The reason for this is to avoid possible bias induced due to the fact that players were already experienced in professional basketball or they were older in age.

Players who were drafted from 1999 up to 2006 were considered and their second year on-court performance metrics statistics were gathered, such as the points scored (PTS), total rebounds (TRB), offensive rebounds (ORB), assists (AST), blocks (BLK), steals (STL), turnovers (TO), and minutes played (MP). Additionally, it considers the efficiency of a player on both offense and defense by looking at percentages of successful field goals (FG%), three-pointers (3P%), and free throws (FT%). Other variables that are included in the model are age (AGE), during the sophomore year, games played (GP) in that season, and draft pick (DP) selection. Furthermore, the player's position was taken into account, with forwards and centers having a higher probability of staying in the NBA longer due to their height and weight as shown from previous research papers.

To provide deeper insights into player efficiency, two new variables were introduced: the assist-to-turnover ratio (AST/TO) and the assist-plus-points-to-turnover ratio ((AST + PTS)/TO). These ratios gauge a player's ability to contribute positively to their team's performance by generating scoring opportunities through assists and points while minimizing turnovers. Higher values of these ratios indicate greater overall contribution and efficiency to the team, potentially prolonging a player's career in the league. By incorporating a wide range of performance indicators and efficiency metrics, our model aims to provide a comprehensive understanding of a player's potential longevity in the NBA, considering individual performance. This multifaceted approach ensures a more accurate assessment of a player's impact and viability over an extended period in the league.

Combining all years of data into a single database for our analysis posed a challenge due to missing information from various sources. This discrepancy arose primarily because some players experienced interruptions in their careers, either due to injuries or leaving the league temporarily during their second year, only to return later. To address this issue, a meticulous approach was adopted to defining a player's sophomore year. Instead of simply relying on the consecutive calendar year, the true second year in the league was identified by considering only those seasons where the player actively participated in at least one game with playing time. This ensured a more accurate representation of each player's progression and continuity within the league. All computations took place using the statistical software *R* Team (2023).

2.1 Descriptive statistics of the data

Table 1 presents the descriptive statistics of the data. Out of the 322 NBA players in our sample, 156 (48.45%) attained the full retirement scheme. Further, only 4 players are still active during the 2023-2024 season, namely LeBron James, Chris Paul, Rudy Gay and Kyle Lowry. It is worthy to mention that James is among the few players who have played in the league for 21 years[3] and Lowry was drafted as the 24th overall pick. Since the performance measures were collected after the second year of the players, it is evident that they have stayed in the league for at least two years and played at least one game during their second year. The majority of the players play solely in the guard or forward positions, while there were players who started one year or even two years after they were drafted.

[3] Vince Carter holds the record with 22 seasons.

All predictor variables are statistically significantly associated with the years in the league as presented in Table 1, and as expected, the players' career longevity is negatively associated with their age and their draft pick. All predictor variables were deemed statistically significant (most p-values were equal to 0.001 or less except for the assist-to-turnover ration which was equal to 0.01) when logistic regression was used for the response variable (attainment of full retirement). The χ^2 test of independence marginally rejected the independence assumptions between the response variable and the draft year (p-value=0.045), between the response variable and the pick round (p-value=0.026) but did not reject it between the response variable and the position they play (p-value = 0.227). Lastly, Welch's F-test did not reject the assumption of equal mean years in the league across the 8 seasons under study (p-value=0.111).

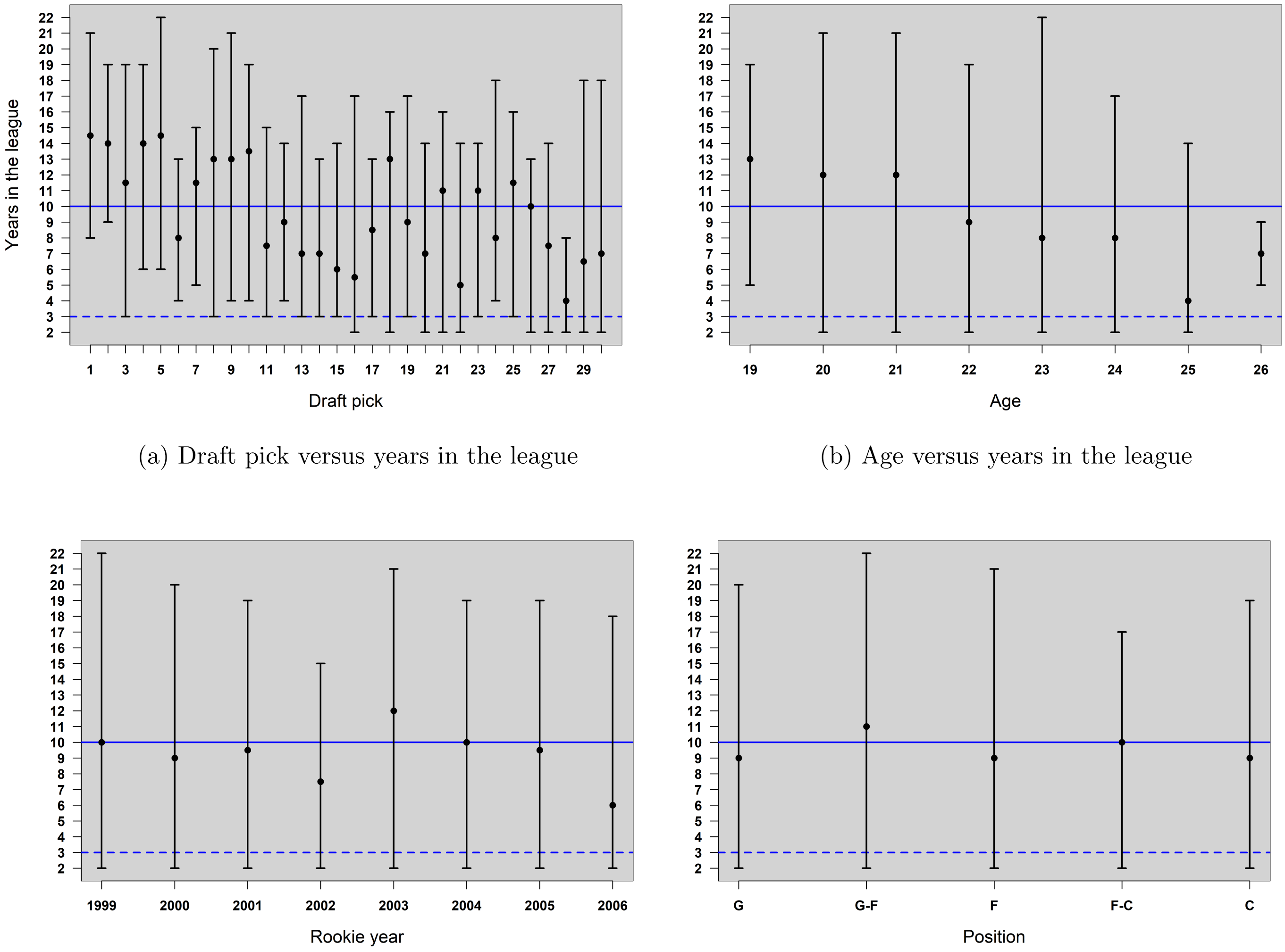
Figure 1 visualizes the effect of the player's age, their first year in the league and their draft pick versus the number of years they survived in the league. Evidently, there is a negative correlation between the age and the number of years, which is statistically significant (p-value<0.001) as depicted in Table 1.

Figures 2 and 3 visualize the estimated distributions of the both groups in the performance statistics under study. In most cases, the distributions of the players who survived in the league for at least 10 years are shifted towards the right compared to those who did not last for that long. Following Psathas et al. (2023) the energy test for the equality of distributions (Székely et al., 2004) of each variable among the two groups of players provided evidence against their equality, for all features (all p-values were equal to 0.001, except for the assists-to-turnover ratio for which the p-value was equal to 0.011). The same conclusion was drawn for the equality of the joint distribution of these 14 variables (p-value=0.001). We further applied the energy test to the joint distributions arising from all possible combinations of the variables and the result was the same, in all cases the hypothesis of equality of distributions was rejected[4].

Table 1: Descriptive statistics of the data. The first column refers to the Pearson correlations between years in the league and the predictor variables, with the statistically significant correlations at the 5% significance level appearing in bold.

Variable	Cor	Min	Max	Median	Mean	Std	Position	Draft year	Year started
YRS		2.000	22.000	9.000	8.991	4.932	G: 105	1999: 75	Same year: 274
AGE	**-0.328**	19.000	26.000	23.000	22.484	1.565	G-F: 38	2000: 34	1 year later: 47
GP	**0.448**	1.000	82.000	63.000	56.304	23.925	F: 104	2001: 40	2 years later: 1
FG%	**0.253**	0.000	0.750	0.437	0.433	0.084	F-C: 35	2002: 30	
3P%	**0.204**	0.000	1.000	0.268	0.217	0.172	C: 40	2003: 31	
FT%	**0.236**	0.000	1.000	0.732	0.7014	0.154		2004: 35	
MP	**0.576**	1.000	42.400	18.000	19.525	9.812		2005: 42	
PTS	**0.579**	0.000	27.200	6.000	7.492	5.131		2006: 35	
TRB	**0.444**	0.000	12.500	2.950	3.413	2.241			
ORB	**0.322**	0.000	4.000	0.800	1.032	0.773			
AST	**0.400**	0.000	9.300	1.000	1.564	1.695			
BLK	**0.283**	0.000	2.600	0.300	0.428	0.459			
STL	**0.488**	0.000	2.200	0.500	0.623	0.432			
TO	**0.508**	0.000	4.200	1.000	1.184	0.753			
ASTO	**0.129**	0.000	6.667	1.000	1.212	0.820			
ASPTTO	**0.169**	0.000	44.000	9.100	9.313	4.105			
DP	**-0.197**								

[4] The same conclusions were reached with the test of equality between the means using the Welch's t-test and the mean vectors using the James test (James, 1954).



(a) Draft pick versus years in the league

(b) Age versus years in the league

(c) Rookie year versus years in the league

(d) Position versus years in the league

Figure 1: Years in the league (minimum, maximum and median) according to (a) draft pick, (b) age, (c) rookie year and (d) position. The blue lines indicates the first and the second limit year to attain the minimum and the full, retirement scheme, respectively.

3 Retirement attainment and identification of the key performance factors

Five different statistical and machine learning algorithms[5] were utilised[6] in order to predict the probability of a player to obtain the full retirement scheme. The following algorithms were used:

- Elastic net (EN) (Zou and Hastie, 2005): This a regularised regression model that combines, linearly, the penalties of LASSO (Tibshirani, 1996) and ridge regression (RR) (Hoerl and Kennard, 1970) and is implemented in the package *glmnet* (Friedman et al., 2010). Logistic regression links the estimated probability of the event y_i (full retirement scheme for the i-th player) to some predictors $X_1, \ldots, X_p$ in a

[5] We ran more algorithms, such as SVM with polynomial kernel and linear kernel, $k-NN$, naive Bayes, but none of them performed better. We further applied ensemble learning of all algorithms, yet there was no improvement, so we decided not to show the results.

[6] The statistical software R (Team, 2023) was employed using the necessary R packages for each algorithm.

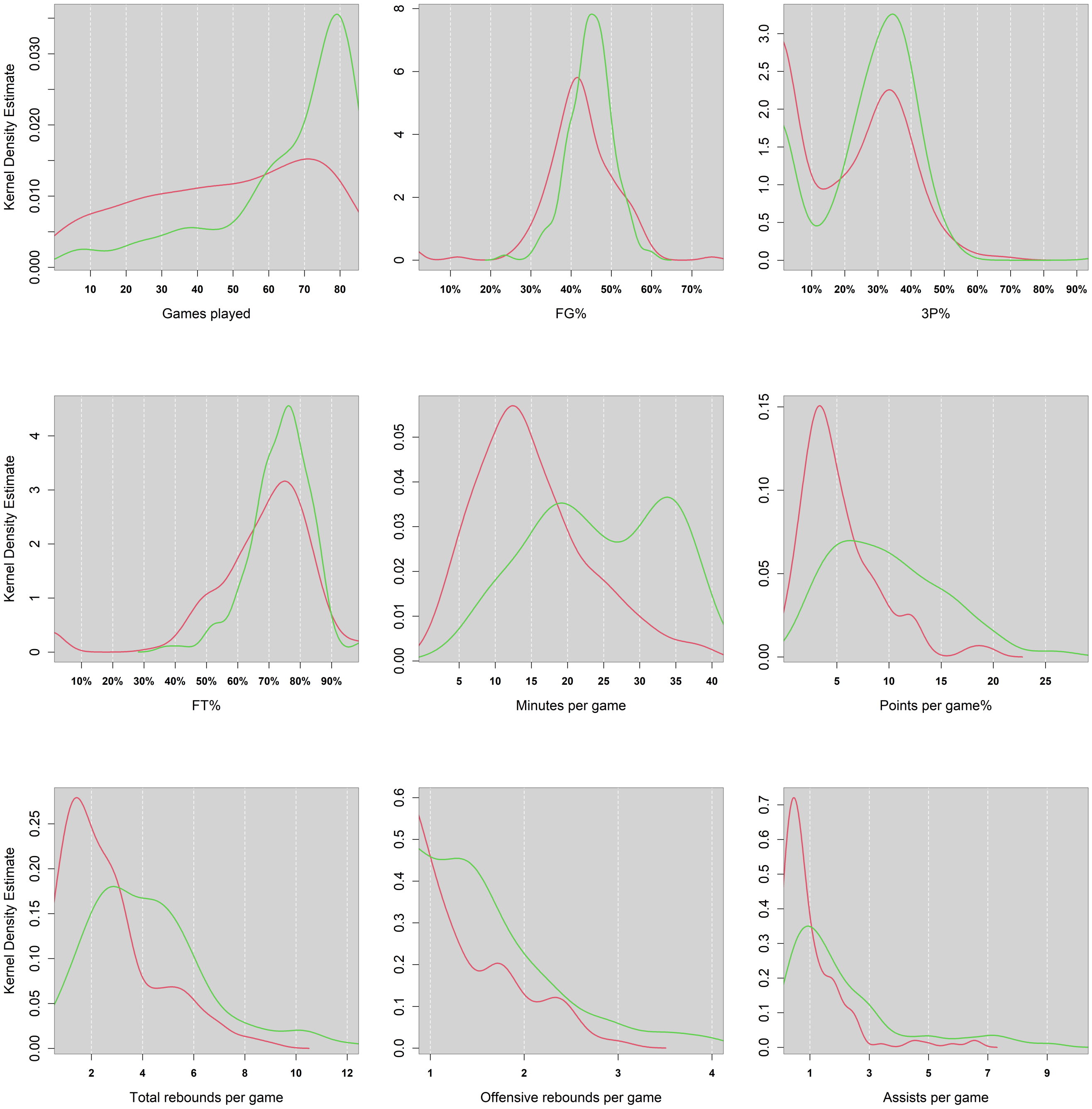


Figure 2: Kernel density estimates for some performance statistics. The red line indicates those who did not survive in the league for then 10 years, whereas the green line corresponds to those who survived in the league for at least 10 years.

non-linear manner

$$P(Y=1|\mathbf{X}) = \frac{e^{\beta_0+\sum_{j=1}^{p}\beta_j X_j}}{1+e^{\beta_0+\sum_{j=1}^{p}\beta_j X_j}}$$

EN estimates the β and β coefficients by minimizing the constrained Kulback-Leibler divergence loss

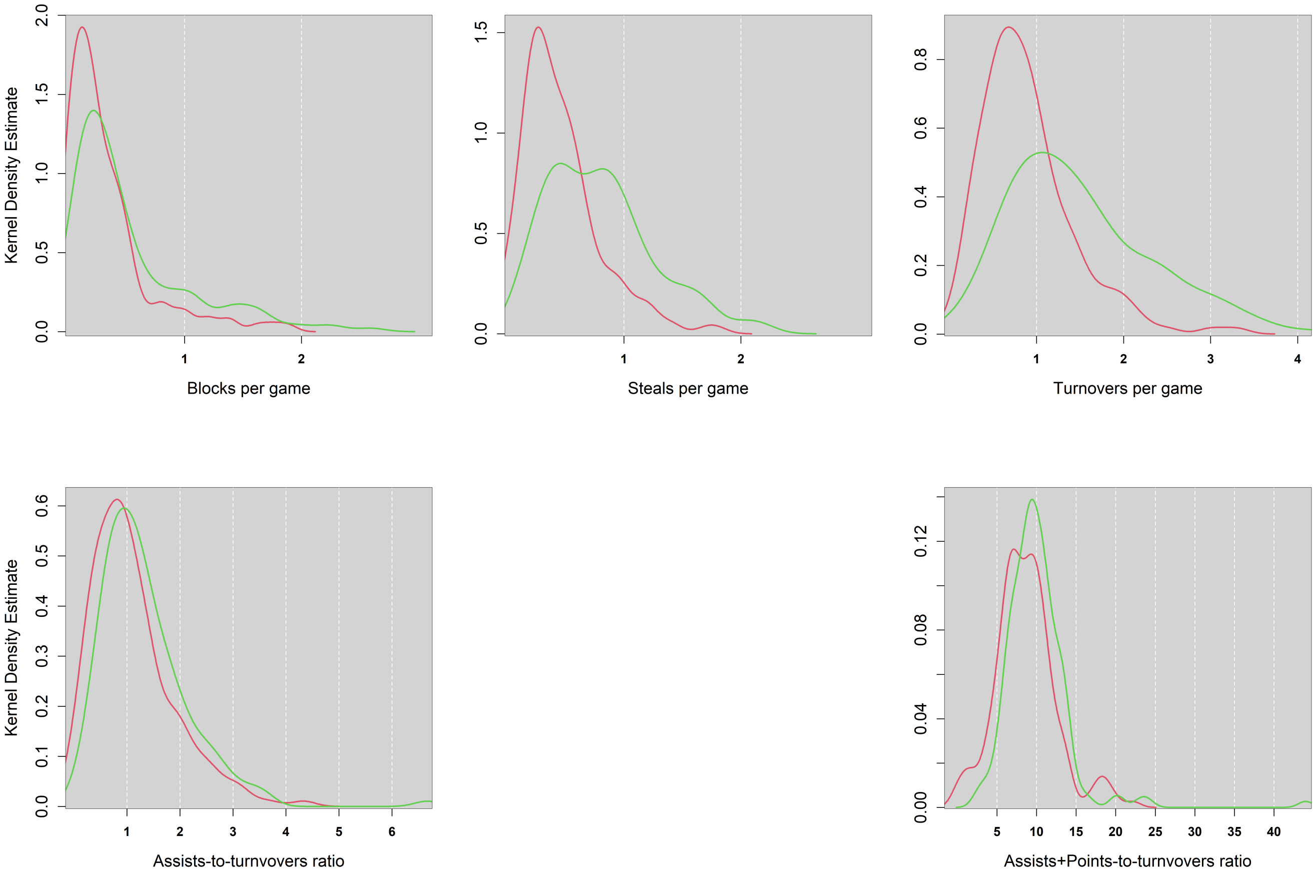


Figure 3: Kernel density estimates for some performance statistics. The red line indicates those who did not survive in the league for then 10 years, whereas the green line corresponds to those who survived in the league for at least 10 years.

function

$$-\sum_{i=1}^{n}\left[y_i\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)-\log\left(1+e^{\beta_0\sum_{j=1}^{p}\beta_j x_{ij}}\right)\right]+\lambda\left[(1-\alpha)\sum_{j=1}^{p}\frac{\beta_j^2}{2}+\alpha\sum_{j=1}^{p}|\beta_j|\right],$$

where $\lambda$is the Lagrangian parameter. The term $\sum_{j=1}^{p}\frac{\beta_j^2}{2}$ refers to the RR imposed constraint, whereas the second term $\sum_{j=1}^{p}|\beta_j|$ refers to the LASO imposed constraint and the α parameter sets the weight allocated to each of the two constraints. If $\alpha = 0$ one ends up with the RR, while if $\alpha = 1$ LASSO emerges.

- Projection pursuit regression (PPR) (Friedman and Stuetzle, 1981): This is a non-parametric smoother and is available as a core function in *R* (Team, 2023). PPR relates the response variable y_i to some predictors via $y_i = \sum_{r=1}^{R} f_r\left(\sum_{j=1}^{p}\beta_j x_{ij}\right) + \epsilon_i$, where the $f_r s$ area collection of R smoothing functions and ϵ_i denotes the error term. The β coefficients are estimated via minimization of the sum of squares of the errors $\sum_{i=1}^{n}\epsilon_i^2 = \sum_{i=1}^{n}\left[y_i \sum_{r=1}^{R} f_r\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)\right]^2$.

- Support vector machines (SVM) (Cortes and Vapnik, 1995): This a non-linear kernel based algorithm, whose implementation in the package *e1071* (Meyer et al., 2023) was employed. In our case, the binary SVM problem is formulated as:
  - $w \cdot x_i + b \geq +1$ for $y_i = +1$ (class 1, full pension scheme),
  - $w \cdot x_i + b \leq -1$ for $y_i = -1$ (class 0, no full pension scheme).

Combining the two equations we obtain $y_i(w \cdot x_i + b) - 1 \geq 0$, for $y_i = +1, -1$[7]. Thus we have two hyperplanes passing through the support vectors, $w \cdot x + b = +1 : H1$ and $w \cdot x + b = -1 : H2$. The SVMs find the hyperplane that separates data by the largest margin $\frac{2}{||w||_2^2}$ and objective is to minimize $\frac{||w||_2^2}{2}$ such that $y_i(w \cdot x + b) - 1 \geq 0$

- Random forest (RF) (Breiman, 2001): Another non-linear, decision trees based, algorithm that is implemented in the package *ranger* (Wright and Ziegler, 2017). RFs independently build multiple decision trees trained on variations of the original data and then combine many models improving the overall results. Each tree is built upon a bootstrap sample of the original data and there is an option to use a randomly selected subset of variables for each tree. In our case we built a RF consisting of 500 trees.

- Gradient boosting machine (GBM) (Friedman, 2001): This is also a non-linear, and generic, algorithm, that iteratively updates the predicted values, and is implemented in the package *gbm* (Greg and Developers, 2024). Starting with some loss function, estimate the fitted values and then obtain some pseudo-residuals which are related to the gradient of the loss function. Use the residuals as the response variable and predict their values using a model. Then update the fitted values by adding, suitably, these predicted pseudo-residuals. The whole process is repeated until the loss function stops improving.

3.1 Methodology

The 10-fold cross-validation (CV) protocol (Hastie et al., 2009) was employed to assess the predictive performance of the algorithms. During the 10-fold CV, the data are split into 10 distinct folds in a stratified manner ensuring that the distribution of the events (full retirement scheme) is retained in these folds. Choose a fold, name it test set and leave it aside. Train an algorithm in the remainder folds (training set) and then use the predictors in the test set to predict the values of the responses (in the test set). Estimate the accuracy of the predictions using some performance metric and repate this process 10 times, untill all folds have played the role of the test set. The area under the curve (AUC) was utilised to measure the predictive performance of the algorithms during the CV protocol. The 10-fold CV was repeated 20 times to account for possible sources of variations among the splits.

Further, variable selection (VS) was performed as an extra step of the analysis using the Boruta non-linear VS algorithm (Kursa et al., 2010) available in the package *Boruta* (Kursa and Rudnicki, 2010). The Boruta algorithm utilizes, iteratively, the RF algorithm to fulfill its purpose (VS) and this allows for computation of the variable importance at each step. This VS procedure and the predictive performance (AUC) of each algorithm were cross-validated, again using the 10-fold CV protocol repeated 20 times.

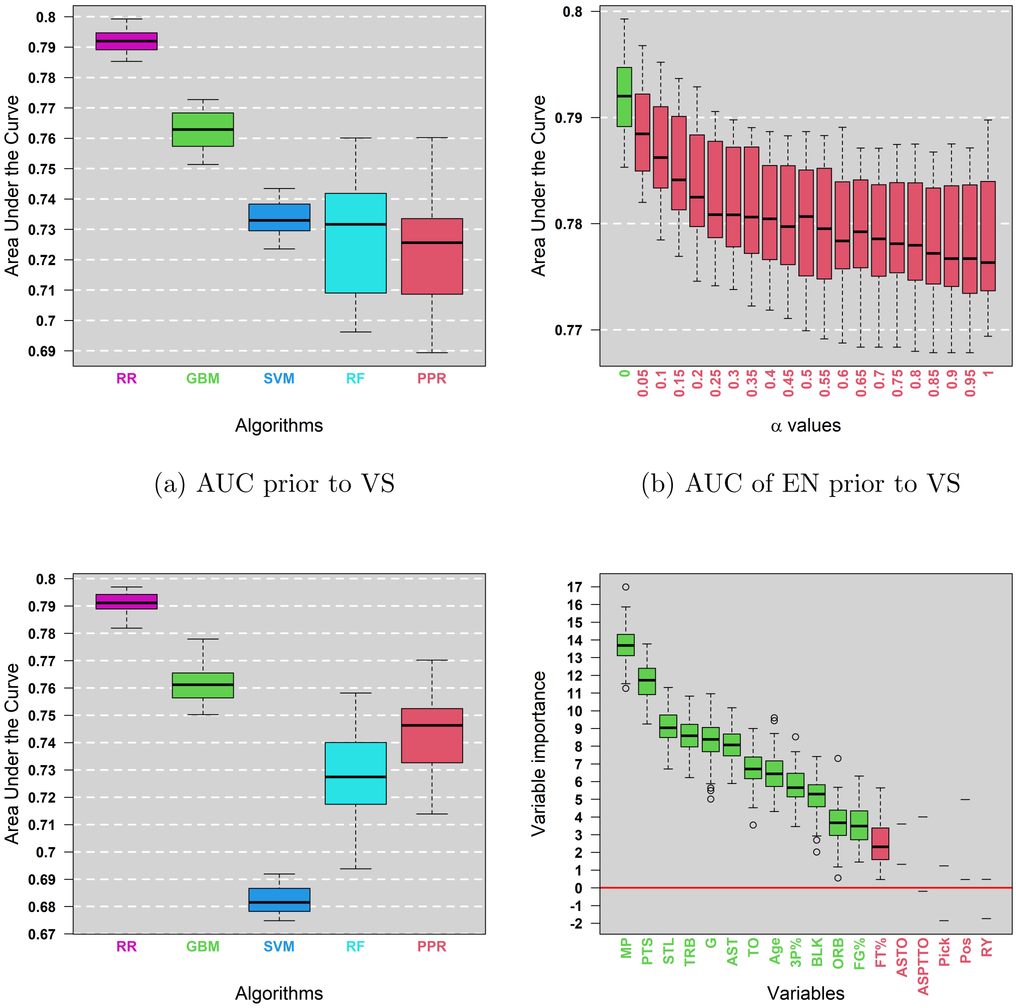
3.2 Results

Figure 4 presents the predictive performance of the 5 algorithms before and after VS, and also the importance of each variable as measured by Boruta. Specifically for EN, as Figure 4(b) shows, the optimal weighting scheme revealed that RR produced the optimal results (before VS). The RR outperformed the other four competing algorithms as shown in Figure 4(a), while Table 2 summarizes the predictive performance of each algorithm before and after the VS procedure. It should be highlighted though that EN was the only algorithm that did not include the position of the players as a predictor variable. Since RR resulted in the optimal results, EN was ran again using feature construction where the squared and the cubic versions of the predictor variables were applied, showing no further improvement of its predictive performance.

Figure 4(c) visualizes the AUC after VS and evidently RR again outperformed the other algorithms, but only this time SVM performed worse compared to prior the VS. Twelve variables were, most of the times, selected by the Boruta algorithm. Out of them, ten were constantly selected, *Age*, *Games played*, *3P%*, *Minutes played*, *Points scored*, *Total rebounds*, *Assists*, *Blocks*, *Steals* and *Turnovers*, whereas *FG* was selected in 96.5%

[7] Note the different coding, instead of 0 and 1 used in logistic regression here -1 and 1, respectively, are used.

and *Offensive rebounds* were selected in the 96.5% and 77.5% of the times, respectively. Boruta was then ran on the whole dataset (no CV was applied) and the variable importance was computed and presented in Figure 4(d) in the order of importance. This figure showcases that the most important variable is the minutes played, followed by the points scored, while the least important ones are the shooting percentages in field goals and the offensive rebounds.



(a) AUC prior to VS

(b) AUC of EN prior to VS

(c) AUC after VS

(d) Variable importance

Figure 4: Box plot of the AUC for (a) each of the five algorithms and (b) of the elastic net for each value of α before the VS. (c) Box plot of the AUC for each of the five algorithms and (d) Variable importance produced by Boruta.

3.2.1 Investigation of the RR model

RR was ran again using the optimal (on average) penalty hyper-parameter selected by the CV protocol and its coefficients were extracted. However, these regression coefficients are biased and thus cannot be used for statistical inference, but can be used though for prediction purposes. The coefficients along with their 95% (bootstrap based) confidence intervals are presented Table 3. As expected, age affects negatively the probability of surviving in the league for at least 10 years, while all other variables affect this probability in a positive manner. The positive sign of the turnovers is also something to expect, since players who tend to possess the ball for longer time they also tend to make more mistakes.

Table 2: Summary statistics of the AUC for each algorithm.

	Prior to VS					After VS				
	RR	GBM	SVM	RF	PPR	RR	GBM	SVM	RF	PPR
Min	0.785	0.751	0.724	0.696	0.689	0.782	0.750	0.675	0.694	0.714
Max	0.799	0.773	0.743	0.760	0.760	0.799	0.778	0.692	0.758	0.770
Median	**0.792**	0.763	0.733	0.732	0.726	**0.791**	0.761	0.682	0.727	0.746
Mean	**0.792**	0.762	0.734	0.728	0.722	**0.791**	0.762	0.682	0.728	0.744

Further, the goodness of fit of the model was assessed. Previously the out-of-sample performance was examined using AUC, but now the in-sample performance of the RR model is assessed using the the Receiver Operating Curve (ROC). Figure 5 displays the ROC curve produced by the fitted values (not the cross-validated predictions) of the RR model. Notably, the in-sample performance is pretty close to the out-of-sample performance, the AUC using the fitted values equals 0.794, whereas the average cross-validated AUC equals 0.791 (see Table 4), providing evidence of no over-fitting.

Finally, Figure 6 contains the Individual Conditional Expectation (ICE) plots Goldstein et al. (2015) that show, as the name reveals, the effect of each variable on the estimated probability of surviving in the league for more than 10 years, conditional on the other variables. The rationale of the ICE plots is the following: Pick a variable X_s and create a new dataset $\mathbf{X}^* = \left\{X_s^i, \mathbf{X}_c\right\}$, where X_s^i denotes the s-th variable whose values contain a single value, the i-th value of that variable. Then feed the dataset $\mathbf{X}^*$ into the RR model, estimate the probabilities of each player staying in the league for more than 10 years and then compute their average. This is the expected probability of player should they have values in the X_s variable equal to the i-th value of this variable. This process is repeated for all $i = 1, \ldots, n$ (where n is the sample size) and for each variable separately.

ICE plots are mostly informative for non-linear models, portraying the non-linear effect of each variable on the estimated outcome and unfortunately they only show the conditional contribution of one variable at the time. In this case though they can provide evidence for the probability of a player attaining the full retirement scheme. For instance, Figure 6(a) shows that on average, players aged 21 years old or less have more than 50% of attaining the full retirement scheme, whereas Figure 6(e) shows that a player should be playing at least 25 minutes per game during their second year should they wish to reach this goal.

Table 3: Coefficients, and their 95% bootstrap based confidence intervals, of the RR model.

Coefficient	Value	95% C.I.
Intercept	-0.3150	(-0.5145, 0.2907)
Age	-0.0441	(-0.0513, -0.0228)
G	0.0031	(0.0027, 0.0042)
FG%	0.4502	(0.3391, 0.7846)
3P%	0.3180	(0.2520, 0.5112)
MP	0.0095	(0.0088, 0.0114)
PTS	0.0172	(0.0158, 0.0213)
TRB	0.0293	(0.0257, 0.0388)
ORB	0.0588	(0.0456, 0.0909)
AST	0.0375	(0.0329, 0.0528)
BLK	0.0929	(0.0713, 0.1552)
STL	0.1839	(0.1656, 0.2443)
TO	0.0995	(0.0900, 0.1276)

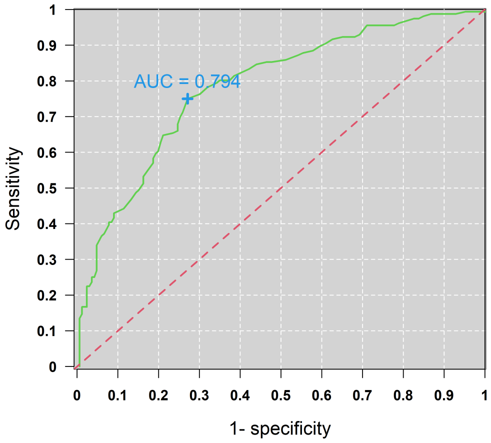


Figure 5: ROC curve of the RR model after VS.

4 Application to the present years

The model's performance was evaluated using more recent data and specifically 77 players who were drafted in 2013 and 2014. Their performance statistics during their sophomore year were again extracted and only the 12 variables of interest were taken into consideration. A notable difference between this data set and the one used in the paper is the proportion of players who survive in the league for at least 10 years. Out of the 322 players drafted between 1999 and 2006, 156 (48.45%) managed to attain the full retirement scheme, whereas in the more recent years, out of the 77 players drafted in 2013 and 2014, only 26 (33.77%) have stayed in league for 10 years or more.

This drop in the percentage can be attributed to many factors. To name a few, the market has expanded worldwide significantly and NBA teams have a wider range of opportunities, outside collegiate players. On the reverse, more players choose to opt out from the NBA because the salaries in the European championships have increased significantly offering competitive contracts.

Differences among the two samples were detected in the performance statistics as well. The energy test of equality of distributions revealed that the distributions of three variables, between the original sample and the recent data, *3P%* (p-value=0.004), *ORB* (p-value=0.002), and *TO* (p-value=0.045) could not be deemed statistically equal. However, the same test showed that the joint distributions of the two groups did not differ statistically significantly (p-value=0.210). The Welch t-test revealed that the means of some more variables, namely the *3P%*, *TRB*, *ORB*, *BLK* and *TO*, were statistically significantly different, and the James test provided evidence that the two mean vectors are statistically significantly different (p-value$<$0.001).

The 12 variable were fed into the RR model, whose coefficients appear on Table 3, yielding an AUC value equal to 0.745. This does not come by surprise for two reasons. At first, the characteristics of the present data may be similar to some degree to those observed in the past data but are not the same and given the rather small sample size, the information contained was not sufficient. This is related to the fact that the NBA is an evolving system that keeps changing, and has become more athletic than in the past. The players used to build the RR were drafted between 1999 and 2006, whereas the validation group contained players who were drafted in 2013 and 2014, exhibiting a maximum of 15 years difference between some players.

5 Conclusions

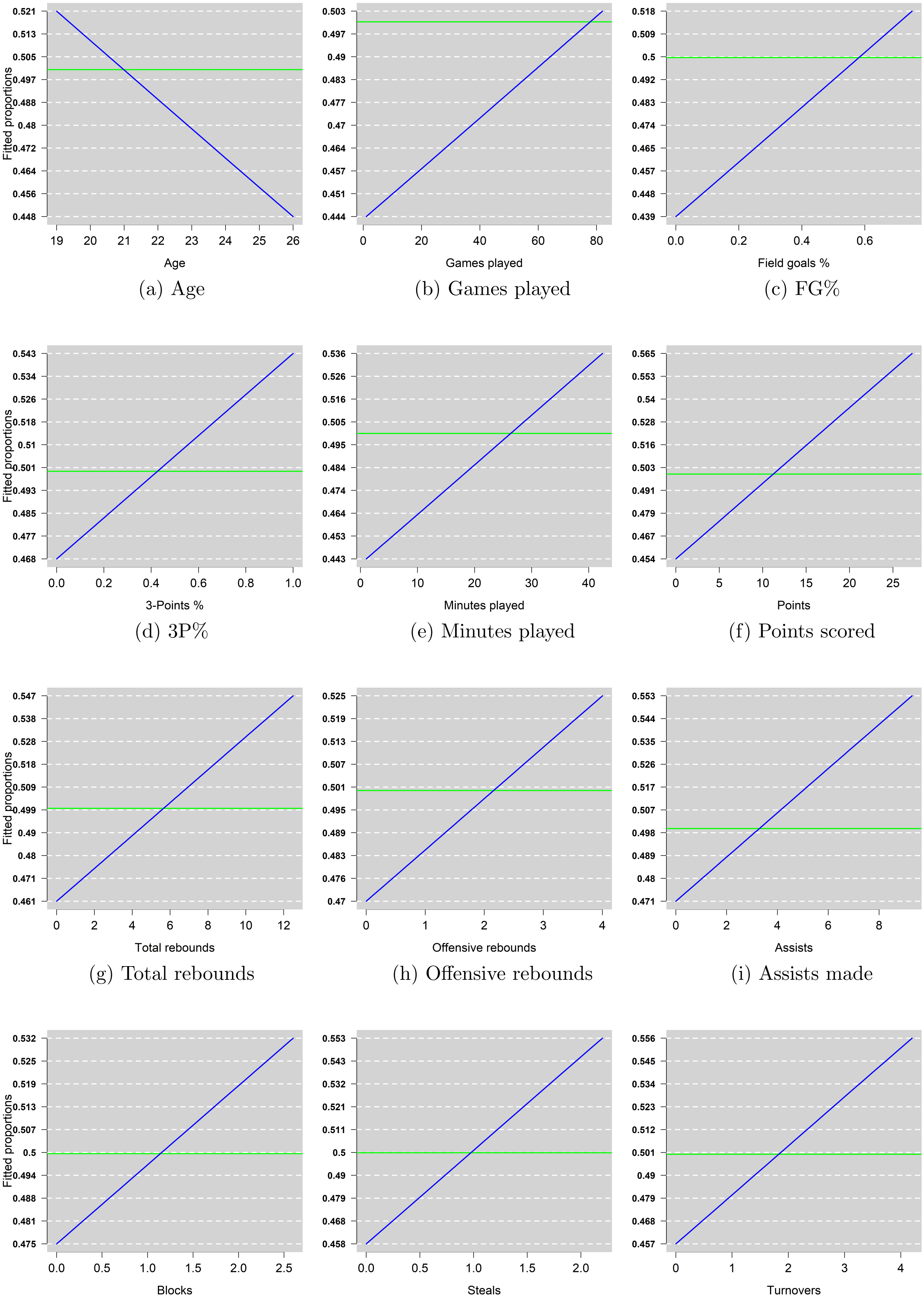
The paper investigated which variables and how accurately they can predict the probability of an NBA player surviving in the league for at least 10 years and thus establish the right to a full retirement scheme. The analysis revealed that out of the 19 performance metrics employed, 12 were deemed statistically important to this end. With the exception of age that evidently has a negative effect on the career duration, all other variables positively affected the probability of surviving for at least 10 years in the league.

Despite using advanced ML algorithms and techniques, we ended up with the ridge logistic regression being the optimal model, in terms of predictive performance. The final model reached a predictive value of AUC equal to 0.791 while the estimated AUC when tested within the training set was equal to 0.794, thus there is evidence to say that we avoided the phenomenon of over-fitting during the repeated 10-fold CV protocol. Using statistics from more recently drafted players not only facilitated an extra performance evaluation of the RR model, but also showed the weaknesses of this analysis. The NBA changes on a yearly basis and as such the model should constantly be updated. A question of interest is to decide on what should the time window be. Should the model use a rolling window and keep only the players who were drafted in the last 15 years, or more years are needed? Further, the analysis included only players who were drafted while in college. How would the inclusion of already professional players (e.g. European) affect the results?

On the economic side, the economic implications of the model are of great importance mainly for the NBA players. Based on their second-year statistics they can compute the likelihood of staying in the league for at least 10 years and can focus on which statistics to improve to increase their chances of securing a full retirement scheme. Investment in skill development by teams is a crucial economic strategy, recognizing the potential for players to have prolonged careers within the league. This proactive approach not only benefits the teams but also motivates players to strategically enhance their capabilities based on insights gained from statistical models. Players can leverage these insights to prioritize skill areas correlated with sustained success, be it refining shooting accuracy, fortifying defensive prowess, or optimizing physical conditioning tailored to their unique strengths and weaknesses, variables that as shown before can have a meaningful impact on career longevity.

Furthermore, players with a statistically higher likelihood of enduring careers (10 or more years) gain significant leverage during contract negotiations. Teams tend to offer more substantial and longer-term contracts to these players, maximizing returns on their investments. Additionally, as has been proven by many accomplished athletes, increased prospects for long-term success in the league often attract lucrative sponsorship and endorsement deals, aligning brands with established athletes. Conversely, players identified as having lower probabilities of longevity may encounter challenges in securing endorsement deals, impacting their potential earnings off the court.

Strategic utilization of statistical models to hone skills not only elevates players' market value during free agency but also fosters a culture of continuous improvement within the league. Teams are more inclined to invest in players demonstrating such commitment and potential for sustained success, thereby fostering increased player mobility and potentially driving up salaries league-wide. Moreover, as the NBA expands its global footprint, statistical models prove instrumental in identifying talent from diverse backgrounds and regions. By adeptly deciphering statistical indicators of success, scouts can effectively unearth and nurture talent from international markets, enriching the league with a more diverse and competitive player pool.



(a) Age

(b) Games played

(c) FG%

(d) 3P%

(e) Minutes played

(f) Points scored

(g) Total rebounds

(h) Offensive rebounds

(i) Assists made

(j) Blocks

(k) Steals

(l) Turnovers

Figure 6: ICE plots of the effect of each predictor variable.

References

Barnes, J. C. (2008). Relationship of selected pre–NBA career variables to NBA players' career longevity. *The Sport Journal*, (April-02).

Breiman, L. (2001). Random Forests. *Machine Learning*, 45:5–32.

Coates, D. and Oguntimein, B. (2010). The Length and Success of NBA Careers: Does College Production Predict Professional Outcomes? *International Journal of Sport Finance*, 5(1).

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823.

Fynn, K. D. and Sonnenschein, M. (2012). An Analysis of the Career Length of Professional Basketball Players. *The Macalester Review*, 2(2).

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65.

Greg, R. and Developers, G. (2024). *gbm: Generalized Boosted Regression Models.* R package version 2.1.9.

Groothuis, P. A. and Hill, J. R. (2004). Exit discrimination in the NBA: A duration analysis of career length. *Economic Inquiry*, 42(2):341–349.

Groothuis, P. A. and Hill, J. R. (2018). Career Duration in the NBA: Do Foreign Players Exit Early? *Journal of Sports Economics*, 19(6):873–883.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

James, G. (1954). Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. *Biometrika*, 41(1/2):19–43.

Johns, W., Walley, K. C., Seedat, R., Thordarson, D. B., Jackson, B., and Gonzalez, T. (2021). Career Outlook and Performance of Professional Athletes After Achilles Tendon Rupture: A Systematic Review. *Foot and Ankle International*, 42(4):495–509.

Kester, B. S., Behery, O. A., Minhas, S. V., and Hsu, W. K. (2017). Athletic performance and career longevity following anterior cruciate ligament reconstruction in the National Basketball Association. *Knee Surgery, Sports Traumatology, Arthroscopy*, 25:3031–3037.

Khalil, L. S., Jildeh, T. R., Tramer, J. S., Abbas, M. J., Hessburg, L., Mehran, N., and Okoroha, K. R. (2020). Effect of Achilles Tendon Rupture on Player Performance and Longevity in National Basketball Association Players. *Orthopaedic Journal of Sports Medicine*, 8(11).

Kursa, M. B., Jankowski, A., and Rudnicki, W. R. (2010). Boruta–a system for feature selection. *Fundamenta Informaticae*, 101(4):271–285.

Kursa, M. B. and Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, 36(11).

Martin, C. L., Arundale, A. J., Kluzek, S., Ferguson, T., Collins, G. S., and Bullock, G. S. (2021). Characterization of Rookie Season Injury and Illness and Career Longevity among National Basketball Association Players. *JAMA Network Open*, 4(10).

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2023). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.* R package version 1.7-13.

Miguel, C. G., Mílan, F. J., Soares, A. L., Quinauad, R. T., Kós, L. D., Palheta, C. E., Mendes, F. G., and Carvalho, H. M. (2019). Modelling the relationship between NBA draft and the career longevity of players using generalized additive models. *Revista de Psicología del Deporte*, 28(3):0065–70.

Petersen, A. M., Jung, W.-S., Yang, J.-S., and Stanley, H. E. (2011). Quantitative and empirical demonstration of the Matthew effect in a study of career longevity. *Proceedings of the National Academy of Sciences*, 108(1):18–23.

Psathas, A., Rallatou, D., and Tsagris, M. (2023). Skin tone of nba players and performance statistics. is there a relationship? *Communications in Statistics: Case Studies, Data Analysis and Applications*, 9(3):234–251.

Staw, B. M. and Hoang, H. (1995). Sunk Costs in the NBA: Why Draft Order Affects Playing Time and Survival in Professional. *Administrative Science Quarterly*, 40(3):474–494.

Székely, G. J., Rizzo, M. L., et al. (2004). Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272.

Team, R. C. (2023). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Wright, M. N. and Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.