Beyond the Pitch: A Data-Driven Approach to Player Careers and Valuations in Football

Constantin-Bogdan Craciun[†]
cocr@itu.dk

Gino Franco Fazzi[†]
gifa@itu.dk

Veron Hoxha[†]
veho@itu.dk

[†]*IT University of Copenhagen*

***Abstract*—High-profile sports like football transcend mere professional disciplines; they are thriving multi-billion-dollar enterprises heavily dependent on diverse factors, with player value standing out as a crucial element. Unraveling the intricacies of how football clubs determine a player's value proves to be a challenging endeavor. Amid the myriad variables influencing a player's market value, one crucial factor is age. This project aims to use survival analysis as a technique to assess the weight of players' age (via their risk of retirement) in their market value. Three distinct models are applied to predict players' market value, incorporating the expected remaining years of their careers as a novel predictor. Surprisingly, the results demonstrate almost negligible improvement in predictive accuracy. We delve into the limitations of our approach, recognizing the nuances in deciphering the complex relationship between age, retirement risk, and market value in football. Additionally, we provide insights into potential avenues for future research, acknowledging the evolving landscape of sports analytics and the continuous quest for a more nuanced understanding of player valuation in the dynamic world of football.**

I. Introduction

High-profile sports transcend mere professional disciplines; they are thriving multi-billion-dollar enterprises heavily dependent on diverse factors, with player value standing out as a crucial element. In the realm of football, the FIFA International Transfer Snapshot recently unveiled a staggering USD 7.36 billion expenditure on transfer fees from June to September 2023[1]. This marks a remarkable 47.2% surge compared to the mid-year period in 2022 and a substantial 26.8% increase over the previous mid-year record set in 2019.

Unraveling the intricacies of how football clubs, both as sellers and buyers, determine a player's value proves to be a challenging endeavor. Numerous variables come into play, making it clear that there is no precise algorithm for accurately establishing a player's cost. Nevertheless, each season, clubs find themselves compelled to estimate the worth of their players, as well as those from other clubs to the best of their ability. Specialized platforms like Transfermarkt leverage the "wisdom of the crowd" principle, akin to the law of large numbers. On Transfermarkt, users engage in discussions on market values through forums, presenting evidence-backed arguments. The platform's staff then evaluate this collective information, combining it with their own research and knowledge to formulate a robust estimate.

Amid the myriad variables influencing a player's market value, one crucial factor is age. The rationale is straightforward: as a player ages, the remaining years of their career diminish, impacting their perceived value in the market. In the following sections we demonstrate how survival analysis can be used as a technique to assess the weight of players' age (via their risk of retirement) in their market value.

A. The dataset

This project's dataset is compiled from two primary sources:

- **Transfermarkt:** a renowned website that aggregates football information, encompassing scores, results, statistics, transfer news, fixtures, and player estimated values;
- and **SoFIFA:** a website offering a comprehensive database of player statistics and information relevant to the FIFA video game series, including player stats and physical attributes.

The final combined dataset comprises 11,159 players from 54 different leagues and 155 different nationalities. A detailed description of the dataset attributes can be found in the Appendix (see Table VII).

II. Methods

A. Histograms, Kernel Density Estimation and Chi-Square Goodness-of-Fit Test

To explore the distribution of age across players in the dataset, we plot the histogram along with its Kernel Density Estimation (KDE) (Figure 1). Kernel smoothing is a non-parametric method to estimate the probability density function of a random variable and allows to compare it to the theoretical normal distribution with same parameters of mean and standard deviation. We use Jacknife resampling to establish the confidence intervals around the mean of the distribution. Jacknife is a form of resampling where given a sample of size n, there can be built an estimator by aggregating the parameter estimates from each subsample of size $(n-1)$ obtained by leaving out one observation at a time. We then employ a Chi-Square Goodness-of-Fit Test [1] to determine how likely it is that the age distribution follows a normal distribution. This test is versatile, applicable to any univariate distribution for which the cumulative distribution function can be calculated.

[1]https://www.fifa.com/legal/media-releases/fifa-international-transfer-snapshot-confirms-record-breaking-figures

Given our binned age data, the chi-square test is well-suited to our observations.

The null hypothesis (H_0) of this test posits that the data conforms to a specified distribution, while the alternative hypothesis (H_a) suggests otherwise. For each age ($i = \{1, ..., k\}$), the chi-square goodness-of-fit statistic (χ^2) is calculated as:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Here, O_i represents the observed frequency for age i and E_i is the expected frequency, computed using:

$$E = N \times \mathcal{N}(a\hat{g}e, s_{age})$$

where $\mathcal{N}(a\hat{g}e, s_{age})$ is the cumulative distribution function for the normal distribution with mean and standard deviation equal to those of the players' ages in the dataset, and N is the sample size.

We set a significance level (α), and if the P-Value associated with the statistic is less than α, we reject the null hypothesis.

Testing for normality is crucial, as many statistical tests rely on specific distributional assumptions, with normality being a common assumption in classical statistical tests.

B. Survival Analysis

While a straightforward approach might link a player's age to their target market value, assuming a linear association, we hypothesize a non-linear relationship between age and the risk of retirement. For instance, the increase in retirement risk from 23 to 24 years is not equivalent to the increase from 34 to 35. Hence, we employ Survival Analysis to estimate and model the risk of a player retiring given their age.

Our application of Survival Analysis, a statistical method widely used in diverse fields such as life expectancy, insurance, and annuities, extends its applicability to model the remaining years of a player's professional career. At the core of Survival Analysis is the "survival function", quantifying the probability that a given event has not occurred by a specified time. In our context, we redefine this as the "Activity function," assessing the probability of a player remaining professionally active beyond a certain age. The focal event is the retirement of a professional player, irrespective of its cause.

Another crucial metric is the "hazard function", depicting the instantaneous failure rate at any given time, considering survival up to that point. By associating a hazard risk with retirement, our objective is to correlate this risk with the current market valuation of players, aiming to replace age by activity rate in the model for valuation forecasting.

We employ in this project various statistical techniques related to Survival Analysis, including Kaplan-Meier [2] curves for estimating the survival function, and Log-Rank tests [4] for comparing survival curves among different groups.

1) Hazard rate, Activity Rate and Kaplan-Meier Estimate: In the context of this project, the hazard rate is defined as **the probability of a player retiring at age i given it has played past age $i-1$**. To operationalize this concept, a set of criteria for player retirement is established (see Section IV-A1). A player is considered *retired* when the following conditions simultaneously apply:

- his last appearance is six months old or older,
- he has no current club,
- he has no current contract,
- he has no current market value.

To calculate the hazard rate, we follow the procedure stated in [3], and let X represent a player's career lifetime. Then f_i is the probability of a player's retiring at age i:

$$f_i = P\{X = i\}$$

In our analysis, the probability of a player still being active after i years of age (A_i) is calculated as:

$$A_i = \sum_{j \geq i} f_j = Pr\{X \geq i\}$$

where f_j is the probability of retiring at age j. Conversely, the *hazard rate*, representing the risk of retirement at age i given that the player has been active until the age $i-1$, is calculated as:

$$h_i = f_i / A_i = Pr\{X = i | X \geq i\}$$

The "conditionality" of the hazard function (conditioned on still being active at a given age) makes it adequate to model the retirement event of a player.

A noteworthy insight mentioned in [3] is that the probability A_{ij} of a player being active past age j given it has been active past age $i-1$ is computed as the product of surviving each intermediate year:

$$A_{ij} = \prod_{k=i}^{j} (1 - h_k) = Pr(X > j | X \geq i)$$

This method allows to estimate the probability of a player being active at a given age without having to wait for his retirement. Further more, our professional life table curves are non-parametric, in the sense that no particular relationship is assumed between the hazard rates.

2) Log-Rank Test: The Log-Rank test [4] assesses whether the Kaplan-Meier curves from two subpopulations are significantly different. In this type of test, the null hypothesis is that the two different curves are equal, and therefore the alternative hypothesis is that the curves differ significantly. It is a non-parametric test based in the χ^2 test. To apply the test, we compute the expected number of retired players at each age for the both subsets separately, given the null hypothesis that the groups have similar curves, and then compute the log-rank test statistic (Z). Asymptotic calculations based on the central limit theorem suggest that $Z \dot{\sim} \mathcal{N}(0, 1)$, allowing us to assess the P-Value under normal conditions.

3) Bootstrap Confidence Intervals: In calculating the Kaplan-Meier estimate for the Activity Rate, we seek a 95% confidence interval (95% CI) around our estimate through bootstrap resampling. Multiple resamples, generated with replacement from our observations, allow us to compute the activity rate for each of those samples, by looking at the values from the resamples that fall in the 2.5%-97.5% range.

Bootstrap resampling provides a non-parametric statistic, eliminating the need to assume normal distribution for our observations or the underlying population. The Central Limit Theorem ensures that the resampling distribution of the effect size converges towards normality.

C. Metropolis-Hastings Monte Carlo Markov Chain

Finally, a complementary strategy was used to assess the certainty of the Kaplan Meier estimates, by use of the Monte Carlo Markov Chain (MCMC) Process [10]. A *Markov chain* is a stochastic process with discrete time parameter, where for each time n, the conditional distributions of all X_{n+j} for $j \geq 1$ given $X_1, \ldots, X_n$ depend only on X_n and not on the earlier states $X_1, \ldots, X_{n-1}$. In symbols, for $n = 1, 2, \ldots$ and for each b and each possible sequence of states $x_1, x_2, \ldots, x_n$, $Pr(X_{n+1} \leq b | X_1 = x_1, \ldots, X_n = x_n)$
$= \Pr(X_{n+1} \leq b | X_n = x_n)$. A Markov chain is called *finite* if and only if there are only a finite number of possible states. The Metropolis-Hastings algorithm [11] is an accept-reject type of algorithm in which a candidate value, θ_{new}, is proposed, and then one decides whether to set $\theta_{(t+1)}$ (the next value of the chain) equal to θ_{new} or to remain at the current value of the chain, $\theta_{(t)}$, which is θ_{cur}. The Metropolis-Hastings MCMC (M-H MCMC) update rule works as follows:

1) Start with a current state (parameter values), θ_{cur}.
2) Propose a new state, θ_{new}, based on a proposal distribution, $q(\theta_{\text{new}}|\theta_{\text{cur}})$.
3) Calculate the acceptance probability, $A(\theta_{\text{new}}, \theta_{\text{cur}})$, which is a function of the likelihoods of the new and current states and the proposal distribution:

$$A(\theta_{\text{new}}, \theta_{\text{cur}}) = \min\left(1, \frac{P(\theta_{\text{new}}|\text{D}) \cdot q(\theta_{\text{cur}}|\theta_{\text{new}})}{P(\theta_{\text{cur}}|\text{D}) \cdot q(\theta_{\text{new}}|\theta_{\text{cur}})}\right)$$

   where $P(\theta|\text{D})$ is the posterior probability of θ given the Data (D), and $(\theta_{\text{new}}|\theta_{\text{cur}})$ is the probability of proposing θ_{new} given θ_{cur} (the proposal distribution).
4) Generate a random number u from a uniform distribution over $[0, 1]$. If $u \leq A(\theta_{\{new\}}, \theta_{\text{cur}})$, accept the new state. Otherwise, retain the current state.
5) Repeat the steps for a large number of iterations to generate a chain of samples.

The implementation of the M-H MCMC has followed the general steps and ideas of the Metropolis MCMC, but has been adapted to incorporate information from the Professional Life Table (see Table III) generated in the Survival Analysis section. The modified version is:

1) **Initialization:** Initialize the algorithm with a random state where each parameter is assigned a random value drawn uniformly between 0 and 1.

TABLE I
Parameters used in H-M MCMC. The paramers differ by the age category.

Parameter	Value
# of Iterations	10,000 - 25,000
Step size	0.01-0.05
Burn-out	$\frac{1}{2}$ # of Iterations.
Target acceptance	0.25

2) **Proposal Generation:** For each iteration, propose a new state by adding Gaussian noise to the current state. The noise is centered at zero with a standard deviation equal to the `step_size`.
3) **Weighted Likelihood and Prior Calculation:**
   - Calculate the weighted likelihood for each proposed parameter. The weights are determined by the ratio of non-retired players to the number of players, reflecting the proportion of the sample that has not retired. These weights effectively act as a prior, adjusting the likelihood based on the sample size.
   - Calculate the weighted prior using the Gaussian probability density function, where the mean is set to the survival probabilities from the data and the standard deviation is a fixed value (0.1 in this case).
4) **Acceptance Ratio and Decision:**
   - Compute the acceptance ratio for each parameter by dividing the product of the new likelihood and new prior by the product of the current likelihood and current prior.
   - Accept the new parameters if the acceptance ratio is greater than a random number drawn from a uniform distribution. If not, retain the current parameters. This determines the next state of the Markov chain.
5) **Adaptive Step Size and Iteration:**
   - During the burn-in period (the first half of the iterations), adjust the `step_size` based on the actual acceptance rate versus the target acceptance rate. This adaptive approach aims to optimize the step size to ensure an efficient exploration of the parameter space.
   - Repeat the process for a large number of iterations, collecting samples to form the Markov chain.

The parameters used were selected using a trial-and-error approach, and can be seen in Table I.

D. Prediction Models for Market Value

This section outlines the statistical and machine learning methodologies applied for predicting the market values of football players. These methods include data preprocessing, handling missing values, feature scaling, and applying regression models to establish a relationship between player attributes and their market values.

Data Preparation: The dataset consists of various player attributes and performance metrics. Missing values in numerical features are handled by median imputation, a strategy that minimizes bias, especially in the presence of outliers.

Feature Standardization: The data undergoes standardization through the subtraction of the mean and division by the standard deviation for each feature. This process, known as z-score normalization, results in features having zero mean and a standard deviation of one, which is essential for models sensitive to variable magnitude.

Model Implementation: The regression models are implemented using Python's scikit-learn[2] library. All models are integrated within pipelines that combine preprocessing (standardization) and the regression algorithm. The selected models are:

1) Lasso Regression: Implements L1 regularization, effectively shrinking some coefficients to zero and thus performing feature selection [12]. This model is particularly useful when we suspect some features to be irrelevant or redundant.

2) Ridge Regression: Implements L2 regularization, penalizing the square of coefficients and distributing the penalty across all features [13]. This model is beneficial in scenarios with high multicollinearity among features, distributing the penalty across all features to maintain model stability.

3) Random Forest Regression: A method using multiple decision trees to improve prediction accuracy and control overfitting [14]. It is configured with specific hyperparameters like the number of trees and maximum tree depth.

For the linear regressors Lasso and Ridge, we force all predictions to be non-negative, by setting negative predictions to zero.

Hyperparameter Tuning: GridSearchCV[3] was used for hyperparameter optimization in all models. The range of alpha values (regularization strength) was chosen to balance between model complexity and overfitting risk. For Lasso regression, the tried alpha values were [0.001, 0.01, 0.1, 1, 10, 50, 100, 200, 500], and for Ridge regression, the tried values were [0.0001, 0.001, 0.01, 0.1, 1, 5, 10, 50, 100]. These values were chosen to explore a broad spectrum of regularization strengths, from very light (close to zero) to very strong, to identify the optimal balance between bias and variance in the model. The optimal alpha values for both models were determined based on their ability to minimize the negative mean squared error during the cross-validation process. This ensured that the chosen models not only balanced complexity and predictive accuracy but also performed consistently across different subsets of the data. For the Random Forest model, the hyperparameters tuned were the number of estimators [50, 100, 150], tree depth [10, None[4]], minimum number of samples required to split a node [2, 5]. The resulting best parameters found for each model can be seen in Table II.

Feature Selection: The features included in the models are chosen based on their relevance and potential impact on a player's market value, ensuring the model's interpretability and effectiveness. Features used for the models can be found in Appendix (see Table VIII).

[2]*scikit-learn Documentation* (link)

[3]*GridSearchCV Documentation* (link)

[4]When set to None, there is no maximum depth of the tree.

TABLE II
Best parameters for selected models found during Grid-Search.

Model	Best parameters
Lasso	$\alpha = 500$
Ridge	$\alpha = 0.1$
Random Forest	Max Depth = None Min. Samples Split = 5 Nr. Estimators = 150

Train-Validation-Test Split: To ensure robust model performance, the dataset is partitioned into training, validation, and testing sets. The training and validation sets play a pivotal role in the Grid Search process (see Section II-D3), accompanied by a 5-fold cross-validation, aimed at identifying optimal parameters for each model. Subsequently, the best model, along with its refined parameters, is selected based on this comprehensive exploration. To assess the model's real-world effectiveness and generalization capability, the hold-out test set is employed, providing a reliable evaluation of the chosen model's predictive prowess.

Performance Metrics: To evaluate and compare model performance, metrics such as **RMSE**, **MAE**, and **R^2** are calculated.

Root Mean Square Error (RMSE): Represents the square root of the average squared differences between predicted and actual market values, offering a measure of prediction accuracy.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

Mean Absolute Error (MAE): Calculates the average absolute differences between predicted and actual values, providing an indication of prediction precision without penalizing large errors excessively.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

R-squared (R^2): Depicts the proportion of variance in the dependent variable that is predictable from the independent variables, serving as a gauge of the model's explanatory power.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean of the actual values, and n is the number of observations.

III. Results

A. Survival Analysis - Student: Gino F. Fazzi

Our initial examination focuses on the distribution of players' ages across the dataset, as depicted in Figure 1, where a

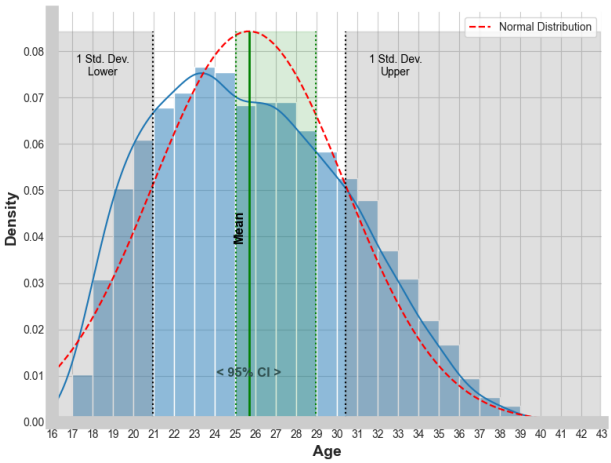

Fig. 1. Histogram depicting the distribution of players age with a kernel density estimate overlaid (solid blue line) and a fitted normal distribution (dashed red line) ($\bar{X} = 25.7; s = 4.7$). The histogram bins capture the empirical frequency of the data, while the kernel density line and the fitted normal distribution visually highlights the deviation from a normal distribution. Includes 95% CI for the mean, calculated with Jacknife resampling, along with 1 Std. Deviation from the mean.

histogram is complemented by a kernel density estimate. The histogram unveils a subtle right-skewed pattern, suggesting a departure from normality. This skewness indicates a preference for younger ages among active players, with a diminishing presence of older players, evident in the right tail of the distribution. To rigorously evaluate normality, a chi-square goodness-of-fit test is employed. The test decisively rejects the null hypothesis (P-value=6.398×10^{-54}), compellingly affirming that the distribution of players' ages does not adhere to a normal distribution.

1) Professional Life Table and Bootstrap Confidence Intervals: The final resulting *professional life table* for all players in our dataset, categorized by years of age, is presented in Table III. Hazard rates are estimated as binomial proportions, reflecting the ratio of retired players to the total players within each age group. The table incorporates 95% confidence intervals, calculated non-parametrically through bootstrapping ($B = 10,000$). The final column ($\hat{e}$) features the anticipated number of remaining active years for players of a given age, derived from the remaining activity function rate.

2) Kaplan-Meier Curve: The Kaplan-Meier (KM) survival curves, showcased in Figure 2, offer a nuanced depiction of players' professional trajectories. Notably, the activity rates trace a compelling narrative—exhibiting a relatively high plateau with a gradual decline for players below 30 years old, indicative of sustained professional engagement during the initial phase of their careers. However, beyond this juncture, the curves embark on a discernibly steeper descent, hinting at an increased risk of retirement as players progress into their thirties and beyond.

A particularly insightful feature of the KM curve is its ability to reveal the median age of player retirement, elegantly marked by the line intersecting the activity rate at 0.50. In our analysis, this median age emerges prominently at 37 years, serving as a valuable reference point for understanding the pivotal stage at which players are equally likely to remain active or retire.

TABLE III
Football players professional life table; at each age, n = number of players, R = number of retired players, $\hat{h}$ = hazard rate $\frac{y}{n}$, $\hat{A}$ = Active probability estimate (with CI lower and upper bounds), $\hat{e}$ = Expected years to retirement.

Age	n	R	$\hat{h}$	$\hat{A}$	0.025	0.975	$\hat{e}$
16	9	0	0.000	1.000	1.000	1.000	21.014
17	115	0	0.000	1.000	1.000	1.000	20.014
18	344	2	0.006	0.994	0.985	1.000	19.014
19	562	4	0.007	0.987	0.976	0.996	18.020
20	680	7	0.010	0.977	0.963	0.989	17.033
21	757	3	0.004	0.973	0.959	0.986	16.056
22	793	4	0.005	0.968	0.953	0.982	15.083
23	855	4	0.005	0.964	0.948	0.978	14.115
24	843	1	0.001	0.962	0.947	0.977	13.151
25	763	4	0.005	0.957	0.941	0.973	12.188
26	771	6	0.008	0.950	0.932	0.966	11.231
27	771	5	0.006	0.944	0.926	0.961	10.281
28	703	9	0.013	0.932	0.912	0.950	9.337
29	651	6	0.009	0.923	0.903	0.943	8.405
30	587	8	0.014	0.911	0.888	0.932	7.482
31	535	23	0.043	0.871	0.846	0.897	6.572
32	413	15	0.036	0.840	0.810	0.869	5.700
33	345	22	0.064	0.786	0.751	0.821	4.860
34	245	18	0.073	0.728	0.686	0.769	4.074
35	186	27	0.145	0.623	0.571	0.673	3.346
36	105	11	0.105	0.557	0.498	0.616	2.723
37	59	7	0.119	0.491	0.423	0.562	2.165
38	40	10	0.250	0.369	0.285	0.453	1.674
39	15	3	0.200	0.295	0.195	0.400	1.306
40	7	1	0.143	0.253	0.136	0.378	1.011
41	3	0	0.000	0.253	0.136	0.379	0.758
42	1	0	0.000	0.253	0.133	0.377	0.505
43	1	0	0.000	0.253	0.135	0.377	0.253

3) Kaplan-Meier Curves across Distinct Player Groups: We investigate differences in retirement risk across two distinct player groupings. In each grouping, we conduct tests for significance in their subgroups, adjusting the significance level with Bonferroni's correction. This correction method is a widely utilized strategy for mitigating the impact of multiple testing. It involves dividing the significance level (α) by the number of tests conducted (N) and using $\frac{\alpha}{N}$ as the corrected significance level. This ensures that the family-wise error rate is constrained to the desired level of α. For all subsequent tests, the applied significance level is $\frac{\alpha}{N} = \frac{0.05}{4} = 0.0125$

a) Body Type: Our dataset contains a specific feature categorizing players based on their physical body type. This classification involves two dimensions: one denoting the body build ('Lean', 'Normal' or 'Stocky') and another segmenting their height (< 170 cm, $170 - 185$ cm, or > 185 cm)[5]. We examine different groups based on body build, specifically categorizing players into 'Lean,' 'Normal,' and 'Stocky.' To discern potential differences in the activity function among groups with distinct body types, we generate separate Kaplan-

[5]We exclude a particular label, 'Unique', since only a few players are in this category.

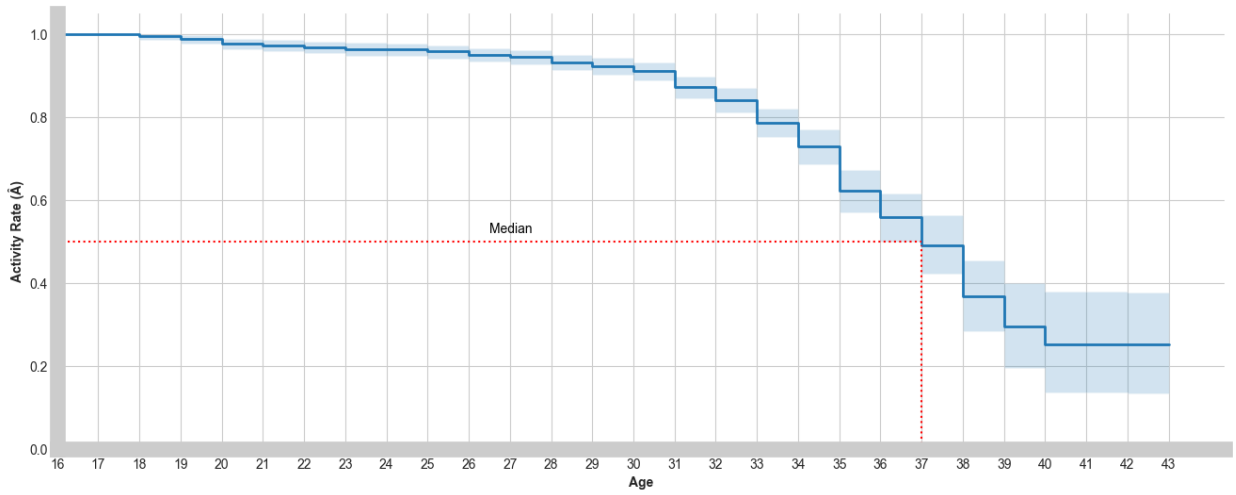

Fig. 2. Activity rate (Â) curve (blue line) by age, along with the 95% confidence interval estimates (shaded blue area).

TABLE IV
Pairwise log-rank test between group of players by Body Type. Includes the χ^2 statistic and the corresponding P-Value. Sample sizes: Lean = 3,632; Normal = 6,747; Stocky = 645. All results are not significant.

Body Type	Stat.	P-Value
Lean vs Normal	0.0110	0.9166
Lean vs Stocky	2.3148	0.1281
Normal vs Stocky	3.2256	0.0725

Meier (KM) curves (see Figure 3). Visual inspection reveals that the 'Lean' and 'Normal' groups exhibit indistinguishable activity rates, while the 'Stocky' group displays a notably lower rate for younger ages until the late 30s, where the curves intersect.

Employing a pairwise Log-Rank Test to assess the significance of these differences (see Table IV), we find that we cannot reject the null hypothesis of equal curves. Consequently, we cannot assert that retirement risk is influenced by a player's body type.

b) Goalkeepers vs Field players: In alignment with prior research findings [6] illustrating the disparate impact of age on market valuation for goalkeepers and field players, we sought to investigate whether these player categories exhibit distinct retirement processes. The Kaplan-Meier (KM) curves for both groups are visually presented in Figure 4.

Upon scrutinizing the KM estimate curves for activity rate between goalkeepers and field players, notable differences become apparent, particularly in ages beyond 30 years. This disparity suggests that goalkeepers may exhibit a propensity to prolong their professional careers compared to field players. One plausible explanation could be that the demands on field players are inherently more intense than those on goalkeepers, potentially allowing goalkeepers to maintain peak performance for an extended duration.

To statistically assess these observed differences, we conducted a log-rank test between the two groups, yielding a significant result (P-Value: 0.0067. Sample sizes: 1,167 goalkeepers and 9,992 field players). This statistical evidence substantiates the notion that goalkeepers and field players indeed experience divergent retirement processes.

B. Metropolis-Hastings Markov Chain Monte Carlo - Student: Constantin-Bogdan Craciun

In order to assess the survival probability of the players, we used an adapted Metropolis-Hastings Markov Chain Monte Carlo (M-H MCMC) that samples from a probability distribution, since the distribution is complex and direct sampling is not feasible. Figures 5 and 6 display two types of plots typically used to analyze the results of MCMC simulations: a trace plot and a posterior distribution histogram. For checking the reliability of the Posterior Distribution of the MCMC, a trace plot is needed in order to visualize a convergence. By visual inspection, one can assess whether the chain has reached equilibrium (stationarity). Convergence is suggested when the trace plot shows a stable horizontal band, indicating that the values are oscillating around a constant mean and the algorithm is no longer exploring drastically different values of the parameter. It is also used in order to asses the Mixing Quality, which shows how well the chain is exploring the parameter space. Good mixing is indicated by a plot that looks like a "fuzzy caterpillar" (see Figure 5).

a) Trace Plots for age 16 and age 41: The trace plots (Figures 5 and 6 - Left) visualize the samples of the survival probability for the age group 16 and 41, as they were drawn over iterations of the MCMC algorithm. The horizontal axis represents the iteration number, and the vertical axis shows the sampled values of the survival probability. The blue line traces the path of the samples, showing how the value of the survival probability changes with each iteration. The dashed green line indicates the mean of the sampled values across all iterations. Ideally, a trace plot should display a "hairy caterpillar" look, which would suggest good mixing and that the chain is exploring the parameter space adequately. In

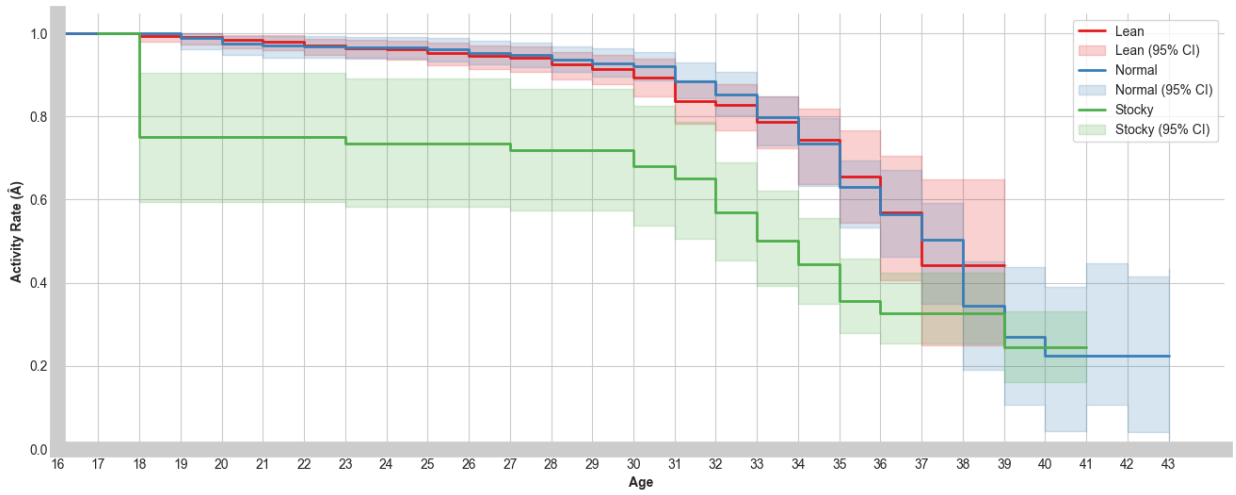

Fig. 3. KM estimate curves for Activity rate (Â) by age, breakdown by body type: Lean (Red), Normal (Blue) and Stocky (Green).

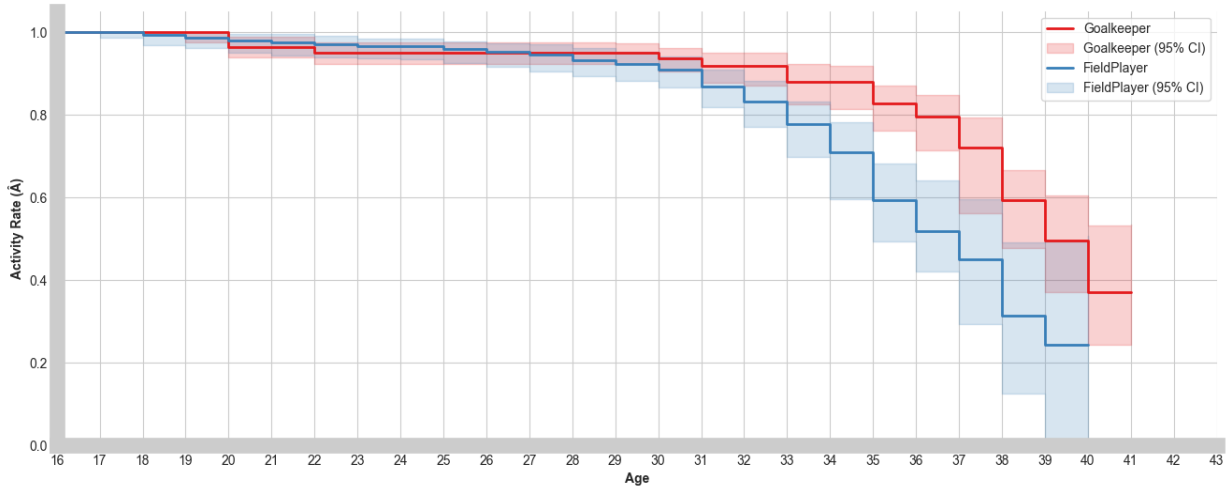

Fig. 4. KM estimate curves for Activity rate (Â) by age, breakdown by player field position: Goalkeepers (Red), Field players (Blue).

the case of age 16 (Figure 5 - Left), the plot shows higher frequency oscillation around the mean, suggesting that the chain may have reached stationarity. For the age 41 (Figure 6 - Left), the flatness of the trace around the mean and the absence of trends or repeated patterns suggest the MCMC algorithm has likely converged.

b) Posterior Distribution for age 16: The histogram (see Figure 5 - Right) shows the posterior distribution of the survival probability for the age group 16 based on the samples drawn during the MCMC procedure. The horizontal axis represents the survival probability, and the vertical axis shows the density or frequency of the sampled values. The shape of the histogram provides an estimate of the posterior distribution of the parameter being sampled—in this case, the survival probability for age 16. The dashed red line marks the mean survival probability as estimated from the MCMC samples. This distribution appears to be right-skewed, with most of the probability mass concentrated around the higher values and a long tail extending towards the lower values. The peak near 1.0 indicates that there's a high likelihood of survival for this age group. The skewness of the histogram implies that while most players at age 16 have a high survival probability, there is uncertainty that spans a range of lower probabilities as well, which is captured by the spread of the histogram.

Overall, these plots are used to assess the performance of the MCMC algorithm and the characteristics of the posterior distribution of the estimated parameter. These plots serve as a second pylon to check the survival analysis.

C. Prediction Models for Market Value - Student: Veron Hoxha

The performance metrics for the best models with three different approaches on the validation set can be seen in Table V with the best resulting parameters found for each model in Table II.

When comparing the performance metrics of all three best models, it is evident that the Random Forest model outstrips

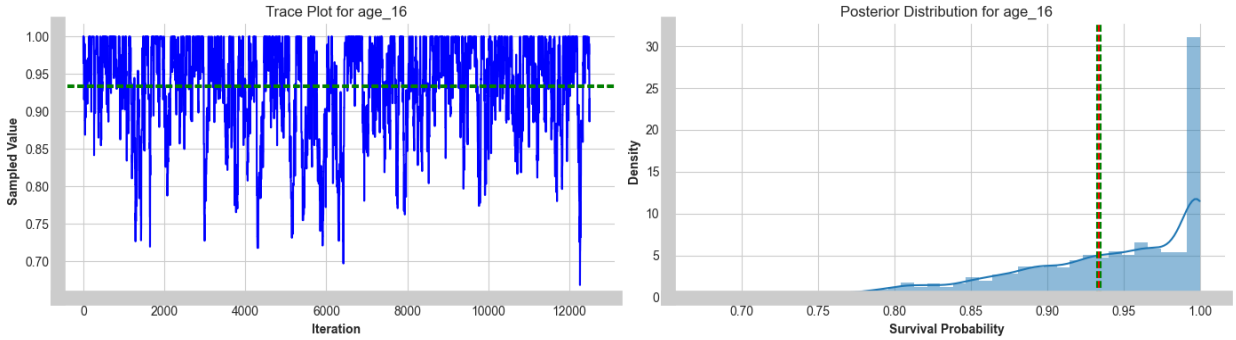

Fig. 5. Trace and Posterior Distribution for age 16.

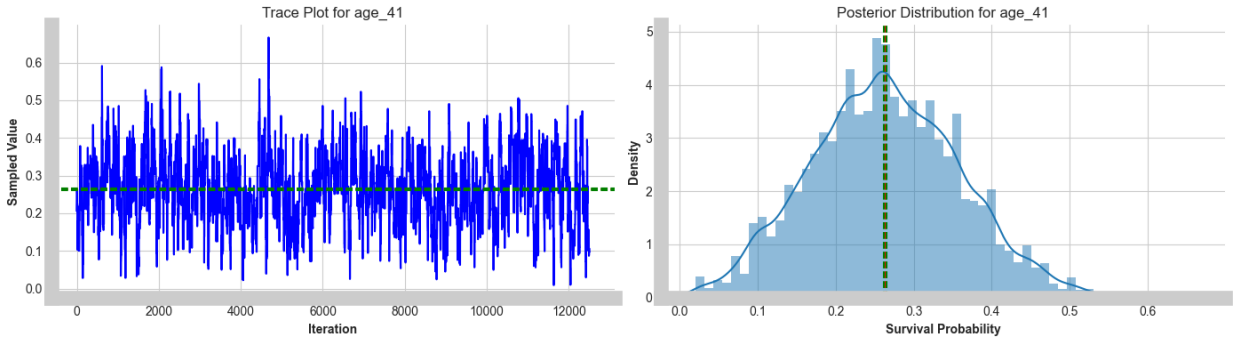

Fig. 6. Trace and Posterior Distribution for age 41.

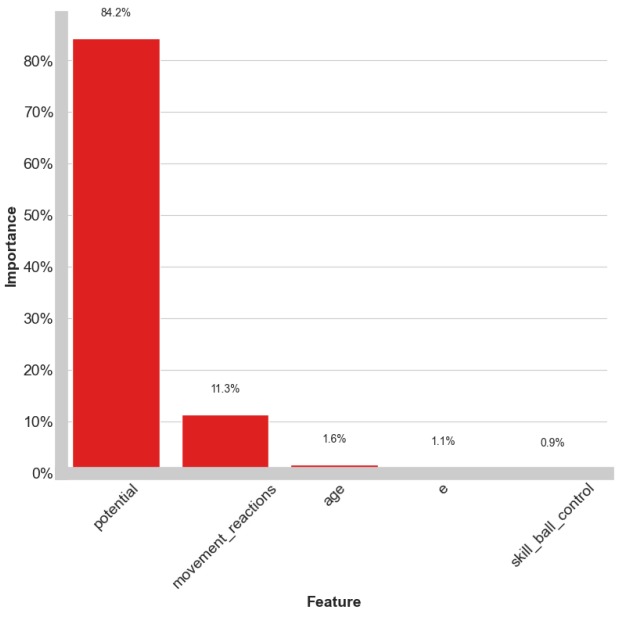

Fig. 7. Random forest top 5 most important features.

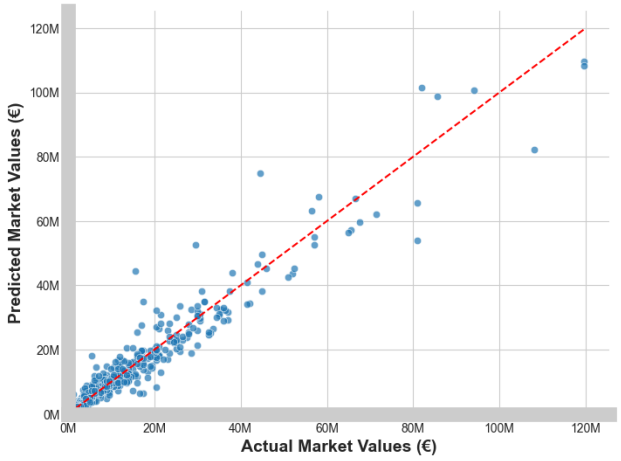

Fig. 8. The Random Forest model's predicted versus actual market values on the test set, demonstrating the model's strong predictive performance.

the Lasso and Ridge regression models significantly in terms of both the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The Random Forest model's RMSE and MAE are markedly lower, indicating a more precise estimation of player market values. Moreover, the coefficient of determination, denoted as R-squared, achieved an impressive peak of **0.922** in the Random Forest model when both the attribute **'age'** and the predictor **estimated remaining playing years** (denoted as $\hat{e}$) are included. This not only underscores the Random Forest's predictive precision but also its ability to utilize these predictors effectively.

In contrast, the Lasso and Ridge regression models exhibit R-squared values of **0.497** and **0.498**, respectively, when incorporating the same attributes. While these values do indicate

TABLE V
Summary of performance metrics for the best regression models.

Model	RMSE (€)	MAE (€)	R-squared (R^2)
w/ Age			
Lasso	6,711,370	3,160,547	0.490
Ridge	6,719,285	3,146,151	0.489
Random Forest	2,658,188	732,401	0.920
w/ $\hat{e}$			
Lasso	6,713,699	3,161,109	0.490
Ridge	6,721,561	3,146,537	0.489
Random Forest	2,692,865	748,439	0.918
w/ Age & $\hat{e}$			
Lasso	6,670,158	3,141,581	0.497
Ridge	6,665,222	3,144,960	0.498
Random Forest	**2,633,141**	**720,389**	**0.922**

a moderate fit, they fall short in comparison to the Random Forest model.

This discrepancy or the superiority of the Random Forest model's performance may be ascribed to a multitude of contributory factors:

- **Complexity and Flexibility:** The model's ensemble approach effectively captures complex, nonlinear relationships within the data, outperforming linear models when interactions are intricate.
- **Robustness to Overfitting:** Despite its detailed modeling capability, Random Forest maintains robustness against overfitting due to its ensemble nature and the averaging of multiple decision trees.
- **Implicit Feature Selection and Interaction:** Random Forest intrinsically selects and utilizes features, including their interactions, enhancing the accuracy without the need for explicit feature engineering.
- **Non-Parametric Nature:** As a non-parametric method, it doesn't assume a particular data distribution, making it suitable for diverse datasets.

By combining the strengths outlined above, the Random Forest model emerges as a well-suited approach for estimating player market values, particularly in datasets where the underlying patterns are not straightforwardly linear.

In our analysis, the Random Forest model emerged as the most precise, with its 5 key features delineated in Figure 7. Notably, the **'potential'** attribute stands out with a significant importance rating of **84.2%**, followed by **movement_reactions**, **age**, **e** and **skill_ball_control**. For an exhaustive evaluation of all features in terms of their importance, please see the comprehensive ranking presented in the Appendix, specifically in Figure 9.

D. Survival Analysis as a predictor - Student: Gino F. Fazzi

In assessing the potential enhancement of our regression models, we investigated the efficacy of replacing the conventional age predictor with the estimated number of remaining years in a player's career. To account for positional differences, we distinguished between goalkeepers and field players when employing this replacement, utilizing the estimated remaining years denoted as $\hat{e}$.

TABLE VI
Sample Predictions from Random Forest Regressor with best parameters. The estimators include age and expected remaining career years ($\hat{e}$)

ID	Name	Actual Val. (€)	Predicted Val. (€)
588171	Faruk Can Genç	700,000	696,780
135747	Luca Lezzerini	1,600,000	1,595,696
135566	Mihai A. Roman	475,000	481,276
41458	Robert Tesche	800,000	806,606
393883	Aleksey Chernov	275,000	281,831
...	...	...	...

Surprisingly, the substitution of age with the estimation of $\hat{e}$ did not yield improvements; in fact, it appeared to adversely affect the performance of the regressors. Undeterred, we explored an alternative approach by incorporating both age and the estimated remaining years $\hat{e}$ into our models. In this scenario, the models demonstrated improvement, albeit marginally, by less than 1%, as detailed in Table V.

The culmination of our efforts is a hypertuned Random Forest model applied to the test data, considering both age and remaining career years ($\hat{e}$).The results showcase a Root Mean Squared Error (RMSE) of 2,184,015.23, a Mean Absolute Error (MAE) of 7,013,56.20, and an R-squared (R^2) of 0.945. This signifies a marked improvement over the validation set. Moreover, the scatter plot in Figure 8 visually reinforces the strength of our best model's predictive performance, providing a clear view of the model's ability to capture the underlying patterns in the data. To provide a tangible representation of our best model predictive capabilities, Table VI presents a snapshot of predicted values on the test data. This excerpt offers a side-by-side comparison of actual market values alongside the corresponding predicted values, offering insights into the model's practical applicability.

IV. Discussion

A substantial body of research has been dedicated to discerning patterns for predicting the market value of athletes, often relying on specific player attributes that signify athletic prowess. In the context of football players, age has consistently emerged as a crucial predictor for market value [5] [6] [7]. This project departs from the conventional use of age as a simplistic feature for market value prediction and introduces a novel approach, exploring retirement risk as a more informative metric through survival analysis methods. To our knowledge, this approach has not been previously attempted.

We employed various techniques, including Kaplan-Meier estimation curves to model the "Activity" function—the probability of a player being active after a given age. We use this activity function to retrieve the expected remaining years of career for players in all ages. Log-Rank tests were utilized to assess if different player groups influenced the risk of retirement at distinct ages, finding that Goalkeepers tend to prolong their career longer than field players. MCMC to have

a second point of view of the survival probabilities of the players, depending on age. Linear regression models, both with and without the activity rate, were implemented to minimize prediction errors in market valuations. Our findings lead us to the conclusion that the remaining estimated years of a player's career, as determined through survival analysis, do not significantly impact the predictive accuracy of market value. While the joint consideration of age and estimated remaining career years ($\hat{e}$) contributes slightly to model improvement, the overall influence of this estimation appears limited in our market valuation predictions. This nuanced insight underscores the intricate nature of factors influencing player market value and reinforces the importance of a multifaceted approach to predictive modeling in the realm of sports analytics.

The comparative analysis of the Ridge, Lasso, and Random Forest regression models has provided significant insights. Both Lasso and Ridge Regression models exhibited similar performance metrics, as shown in Table V. This similarity can be attributed to several factors:

- **Similar Feature Scales**: The dataset's features used for modeling possess similar scales, reducing the impact of standardization and leading both models to perform similarly.
- **Dataset Characteristics**: The inherent nature of the dataset might not distinctly favor the specific strengths of either Lasso or Ridge, leading to convergent performance.

In contrast, the Random Forest model showed superior performance, as explained previously in Section III-C.

A. Limitations

1) Retirement event: The foundation of our survival function centers around the nuanced assessment of player retirement. While this event is often objectively defined in various fields, its objectivity is contextual in our study. Player retirement, induced by the data, doesn't necessarily equate to a definitive retirement, as some players may be inactive without formally retiring or could potentially come out of retirement. The assumptions used to distinguish retired players (see Section II-B1), though valuable, may have limitations affecting accuracy. Furthermore, previously retired players may be missing completely from our data, underscoring the hazard function. In conclusion, the dynamic nature of a player's career introduces intricacies that our analysis may not fully capture.

2) Hazard rates over time: Hazard rates, representing the instantaneous risk of an event, are subject to potential evolution influenced by factors such as medical advances, athletic lifestyle understanding, and technological advancements. The dynamic nature of hazard rates associated with player retirement may exhibit fluctuations over time due to these influences. This evolving landscape adds complexity to retirement risk prediction. The variability in these factors underscores the challenge of relying solely on hazard rates for precise long-term predictions, emphasizing the need for continuous reassessment and adaptation in sports analytics.

3) Negative Market Values in Lasso and Ridge Regression: In our analysis, Lasso and Ridge regression models occasionally predicted negative market values for football players. Since negative values are not feasible in the real-world market context, we adjusted these predictions to zero, as mentioned in Section II-D. This approach, while practical, might underscore the model's performance.

4) MCMC tuning: MCMC algorithms often require careful tuning of parameters (like the proposal distribution in Metropolis-Hastings). Improper tuning can lead to inefficient sampling. In our case, the histogram for age 16 stacks up against the value of 1.00 might suggest a border issue, where the parameter cannot exceed this value due to the nature of probabilities. Therefore the model specification must be extended to ensure it can explore the entire parameter space properly.

B. Future research

a) Injury history: An influential factor in athletes considering retirement is the occurrence of injuries preventing them from performing at a high level. Although not incorporated into our dataset, we recognize the potential significance of injury history in enriching our analysis. Previous research has explored the consequences of specific types of injuries in professional athletes' careers [8], noting that certain injuries can sharply shorten expected professional life. Additionally, studies [9] have investigated the impact of various types of sport participation on the risk of injury, a risk that subsequently cascades to retirement risk, thus proving crucial to our overall analysis.

Future research could aim to incorporate the injury history of a player to improve the survival analysis method and yield a more precise approach to the remaining professional life for players. This integration would contribute valuable insights into the role of injuries in shaping retirement patterns, allowing for a more comprehensive understanding of the intricate relationship between injuries, player longevity, and career trajectories in professional sports.

b) Complex models: Future research could try to implement more sophisticated models, such as those that go beyond conventional regression models. Techniques such as deep learning, ensemble methods, or neural networks present promising avenues for more elaborated analyses, allowing for the incorporation of more complex patterns, non-linear relationships, and intricate interactions within the data. Additionally, the consideration of temporal dynamics and evolving trends could further enrich the predictive power of models. Time-series analyses or recurrent neural networks may offer valuable insights into how market values fluctuate over time, considering external factors such as evolving player performance, changes in team dynamics, or shifts in the broader sports landscape.

REFERENCES

[1] W. J. Ridgman, "Statistical Methods, 8th edn, by G. W. Snedecor & W. G. Cochran. xx 503 pp. Ames: Iowa State University Press (1989).

ISBN 0 8138 1561 6.", The Journal of Agricultural Science, vol. 115, no. 1, pp. 153–153, 1990. doi: 10.1017/S0021859600074104

[2] Kaplan E. L., Meier P., "Nonparametric Estimation from Incomplete Observations", Journal of the American Statistical Association, volume 53, number 282, pages 457-481, year 1958. doi: 10.1080/01621459.1958.10501452.

[3] Efron B., Hastie T., "Computer Age Statistical Inference: Algorithms, Evidence, and Data Science". Cambridge: Cambridge University Press, 2016. doi: 10.1017/CBO9781316576533.

[4] Cochran W. G., "Some Methods for Strengthening the Common χ^2 Tests", Biometrics, vol. 10, no. 4, 1954, pp. 417–51. JSTOR. doi: 10.2307/3001616.

[5] V. B. Jishnu, P. V. H. Narayanan, S. Aanand and P. T. Joy, "Football Player Transfer Value Prediction Using Advanced Statistics and FIFA 22 Data", 2022 IEEE 19th India Council International Conference (INDICON), Kochi, India, 2022, pp. 1-6, doi: 10.1109/INDICON56171.2022.10040117.

[6] Romann M., Javet M., Cobley S., Born D-P. "How Relative Age Effects Associate with Football Players' Market Values: Indicators of Losing Talent and Wasting Money". Sports. 2021; 9(7):99. doi: 10.3390/sports9070099.

[7] Hansoo L., Bayu A. T., Meeyoung C. "Prediction of Football Player Value using Bayesian Ensemble Approach". Journal of Expert Systems with Applications. 2022. doi: 10.48550/arXiv.2206.13246.

[8] Kester B. et al. "Effect of Shoulder Stabilization on Career Length and Performance in National Basketball Association Athletes." Bulletin of the Hospital for Joint Disease (2013) vol. 77,4 (2019): 223-229. doi: 10.1177/23259671231202973.

[9] Malisoux L., Frisch A., Urhausen A., Seil R., Theisen D. "Monitoring of sport participation and injury risk in young athletes", Journal of Science and Medicine in Sport, Volume 16, Issue 6, 2013, Pages 504-508, ISSN 1440-2440. doi: 10.1016/j.jsams.2013.01.008.

[10] DeGroot M. H., Schervish M. J. (2002). "Probability and Statistics". Addison Wesley. Page 188-189. ISBN 10: 0-321-50046-6.

[11] Tobias J.L., "Primer on the Use of Bayesian Methods in Health Economics", Encyclopedia of Health Economics, Elsevier, 2014, Pages 146-154, ISBN 9780123756794. doi: 10.1016/B978-0-12-375678-7.00708-2.

[12] R. Tibshirani, "Regression Shrinkage and Selection via The Lasso: A Retrospective", Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 73, no. 3, pp. 273–282, 2021. doi: 10.1111/rssb.12353

[13] T. Hastie, R. Tibshirani, J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", 2nd ed., Springer, 2009, pp. 43-99. ISBN: 978-0-387-84857-0

[14] T. Hastie, R. Tibshirani, J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", 2nd ed., Springer, 2009, pp. 587-604. ISBN: 978-0-387-84857-0

APPENDIX

TABLE VII
Full dataset features

Source	Feature	Description
Transfermarkt	player_id	Numerical ID.
	first_name	Player's first name.
	last_name	Player's last name.
	last_season	Player's last season appearance.
	country_of_birth	Player's country of birth.
	city_of_birth	Player's city of birth.
	country_of_citizenship	Player's country of citizenship.
	date_of_birth	Date of birth in format yyyy-mm-dd.
	sub_position	Player's sub-position in the field (e.g., Centre-Back).
	position	Player's position in the field (e.g., Defender).
	foot	Preferred foot (e.g., left).
	height_in_cm	Height in cm.
	market_value_in_eur	Market value in Euros.
	highest_market_value_in_eur	Highest historical market value in Euros.
	contract_expiration_date	Contract expiration timestamp in format yyyy-mm-dd HH-MM-SS.
	current_club_domestic_competition_id	Player's current club leage ID.
	current_club_name	Player's current club league name.
SoFIFA	short_name	Player's short name, as displayed in the video game (e.g., L. Messi)
	long_name	Player's full name (e.g., Lionel Andrés Messi Cuccittini)
	player_positions	Player's positions in the field, encoded (e.g., RW, ST, CF)
	overall	Weighted average of ability attributes (scale 0-99).
	potential	Player's potential of the current season (overall rating + value considering age) (scale 0-99).
	value_eur	Market value in Euros.
	wage_eur	Wage in Euros.
	age	Player's age.
	dob	Date of Birth.
	height_cm	Height in cm.
	weight_kg	Weight in Kg.
	club_name	Current club name.
	league_name	Current league name.
	league_level	Current league level (1st to 5th divisions).
	club_contract_valid_until	Current club expiration date.
	nationality_name	Nationality.
	nation_position	Position played in the National team (if applicable).
	preferred_foot	Preferred foot.
	weak_foot	Weak foot.
	skill_moves	Number of special skills available (ordinal, 1 to 5).
	international_reputation	The individual international reputation (ordinal, 1 to 5).
	work_rate	The rate of a player's behavior on the pitch in defending/attacking work (ordinal: Low, Medium, High. E.g, Low/Medium).
	body_type	Body type descrived as body build and height (e.g., Lean (175+)).
	release_clause_eur	Release clause in Euros.
	player_traits	List of different traits associated to the player (Categorical. E.g., "Solid Player, Finesse Shot").
	Player's athletic attributes	All of the following attributes are in the scale 1-99: pace, shooting, passing, dribbling, defending, physic, attacking_crossing, attacking_finishing, attacking_heading_accuracy, attacking_short_passing, attacking_volleys, skill_dribbling, skill_curve, skill_fk_accuracy, skill_long_passing, skill_ball_control, movement_acceleration, movement_sprint_speed, movement_agility, movement_reactions, movement_balance, power_shot_power, power_jumping, power_stamina, power_strength, power_long_shots, mentality_aggression, mentality_interceptions, mentality_positioning, mentality_vision, mentality_penalties, mentality_composure, defending_marking_awareness, defending_standing_tackle, defending_sliding_tackle, goalkeeping_diving, goalkeeping_handling, goalkeeping_kicking, goalkeeping_positioning, goalkeeping_reflexes, goalkeeping_speed, ls, st, rs, lw, lf, cf, rf, rw, lam, cam, ram, lm, lcm, cm, rcm, rm, lwb, ldm, cdm, rdm, rwb, lb, lcb, cb, rcb, rb, gk

TABLE VIII
Feature Columns for the Prediction Model.

Feature Column	**Description**
age	Player's age
$\hat{e}$	Estimated number of remaining years in a player's career, discriminated for goalkeepers and field players
potential	Player's potential rating
skill_moves	Skill moves attribute
pace	Pace attribute
shooting	Shooting attribute
passing	Passing attribute
dribbling	Dribbling attribute
defending	Defending attribute
physic	Physical attribute
skill_dribbling	Dribbling skill attribute
skill_curve	Curve skill attribute
skill_fk_accuracy	Free-kick accuracy attribute
skill_long_passing	Long passing skill attribute
skill_ball_control	Ball control skill attribute
movement_acceleration	Acceleration attribute
movement_sprint_speed	Sprint speed attribute
movement_agility	Agility attribute
movement_reactions	Reactions attribute
movement_balance	Balance attribute
power_shot_power	Shot power attribute
power_jumping	Jumping power attribute
power_stamina	Stamina attribute
power_strength	Strength attribute
power_long_shots	Long shots power attribute
mentality_aggression	Aggression mentality attribute
mentality_interceptions	Interceptions mentality attribute
mentality_positioning	Positioning mentality attribute
mentality_vision	Vision mentality attribute
mentality_penalties	Penalties mentality attribute
mentality_composure	Composure mentality attribute

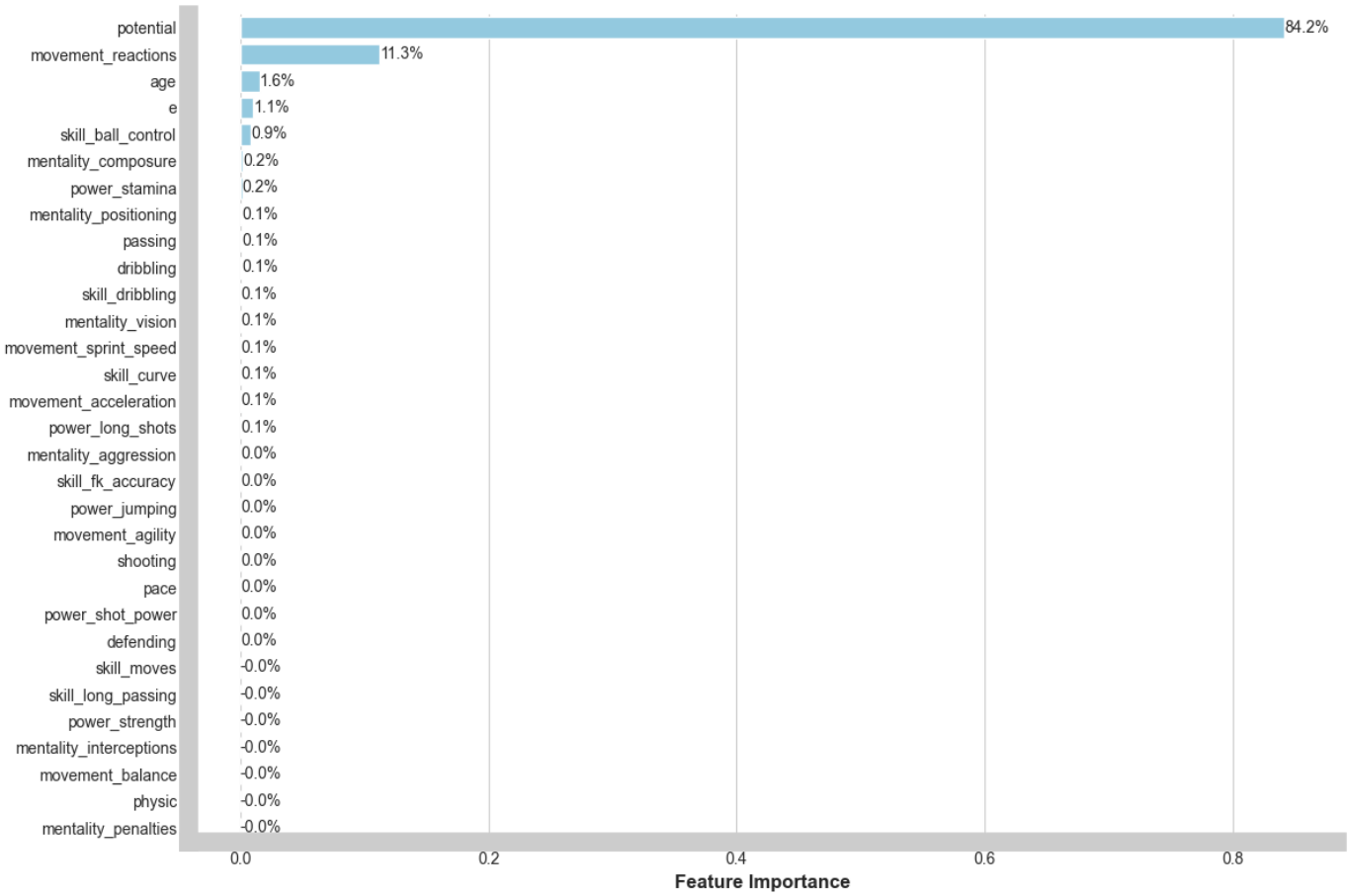

Fig. 9. Random Forest - Feature importance.