Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Multi-Algorithm Machine Learning Framework for Predicting Career Longevity in Professional Basketball: Integrating Draft Combine Measurements and Early-Career Performance

Ran Liu
[1]Shandong Jianzhu University
Yongyan Gao
[1]Shandong Jianzhu University
Yan Zhuang
Qingdao Petrochemical Senior Technical School
Luyu Wang
131530381920163.com

Shandong Sport University

Article

Keywords: machine learning, career longevity prediction, NBA, sports analytics, LIME interpretability, survival analysis, feature selection, class imbalance

Posted Date: May 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9331494/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Multi-Algorithm Machine Learning Framework for Predicting Career Longevity in Professional Basketball: Integrating Draft Combine Measurements and Early-Career Performance

Ran Liu[1], Yongyan Gao[1], Yan Zhuang[3] & Luyu Wang[2]*

[1]Shandong Jianzhu University, Jinan, China. [2]Shandong Sport University, Jinan, China. [3]Qingdao Petrochemical Senior Technical School, Qingdao, China. *Corresponding author: Luyu Wang (13153038192@163.com)

Abstract

This study develops a multi-algorithm machine learning framework to predict whether NBA players will achieve a career of ten or more seasons—the threshold for full retirement benefits. Integrating Draft Combine measurements, draft information, and early-career statistics (first three seasons) for 576 retired players (2000–2015 drafts), we applied Boruta feature selection (28 → 18 features) and trained four classifiers with SMOTE inside ten-fold cross-validation and Optuna optimization. L1-regularized logistic regression achieved the best performance (AUC = 0.840), outperforming Ridge regression (0.829), XGBoost (0.816), and SVM (0.804). LIME analysis revealed that early-career performance contributed 85.2% of the predictive signal, with rebounds and peak scoring as the leading predictors. Supplementary survival analyses using Cox regression (C-index = 0.811) and Random Survival Forests (C-index = 0.768) on the full sample ($N = 651$, including 75 active players) corroborated these findings. The results demonstrate that regularized linear models consistently outperform complex alternatives for this task, and provide actionable insight for player evaluation in professional basketball.

Keywords: machine learning, career longevity prediction, NBA, sports analytics, LIME interpretability, survival analysis, feature selection, class imbalance

Introduction

In modern professional sports, the National Basketball Association (NBA) represents not only the pinnacle of competitive basketball but also a highly commercialized industry of considerable economic scale. As franchise valuations surge and player compensation continues to rise, the accurate assessment of player potential and the prediction of career trajectories have become central concerns for team management [1]. Under the framework of the NBA's Collective

Bargaining Agreement (CBA), a player's career duration directly affects both the franchise's return on investment and the player's post-career welfare provisions. Notably, players who accrue at least ten years of service qualify for the league's full retirement benefit, rendering this milestone a matter of significant institutional and financial consequence [2].

Predicting whether a given player will sustain a career of ten or more years in the NBA remains a formidable challenge. Traditional scouting practices rely heavily on subjective judgment and experiential intuition, whereas the advent of sports analytics has prompted an increasing reliance on objective, data-driven evaluation [3]. The NBA Draft Combine provides standardized measurements of physical attributes and athletic ability, while early-career on-court statistics capture a player's capacity to adapt to the professional game. The effective fusion of these heterogeneous data sources into a high-accuracy predictive model constitutes an important open problem in sports analytics.

Machine learning (ML) methods have gained considerable traction in sports analytics in recent years, spanning applications from game outcome prediction to injury risk assessment, and have demonstrated potential advantages over conventional statistical approaches [4]. Although prior studies have examined the determinants of NBA career longevity, most have relied on traditional survival analysis or simple linear regression [5][6], and have typically drawn on a single data source—either pre-draft measurements or on-court performance statistics, but not both. Moreover, two persistent methodological challenges in this domain—class imbalance (players with careers exceeding ten years constitute a minority class) and insufficient model interpretability—remain inadequately addressed [7].

To address these gaps, the present study develops a multi-algorithm ML framework that integrates pre-draft physical measurements, draft information, and early-career performance statistics to predict whether an NBA player will achieve a career of ten or more seasons. The contributions of this work are fourfold: (1) we assemble a multi-dimensional player dataset and apply the Boruta algorithm for rigorous, statistically grounded feature selection; (2) we integrate SMOTE within the cross-validation pipeline, thereby mitigating class imbalance while preventing data leakage; (3) we evaluate four ML algorithms—L1-regularized logistic regression, Ridge logistic regression, XGBoost, and SVM—and employ LIME to provide both local and global interpretability for the best-performing model; and (4) we conduct supplementary survival analyses using Cox proportional hazards regression and Random Survival Forests to validate the classification findings and to accommodate right-censored observations from active players.

The findings of this study are intended to provide data-driven decision support for NBA franchises with respect to player drafting and development, and to offer methodological guidance for the application and interpretation of ML models within the broader field of sports analytics.

Related Work

Machine Learning in Sports Analytics

Sports analytics, as an interdisciplinary field, has advanced rapidly in recent years, driven by the growth of data science and computational resources. Sarlis and Tjortjis [3] noted that advanced statistical indicators and data mining techniques have become standard tools for evaluating player and team performance in the NBA and European basketball leagues. As data accumulate, ML algorithms have shown considerable promise in predicting game outcomes, assessing player value, and managing injury risk. Bunker and Susnjak [4] conducted a systematic review of ML applications in team sports prediction, observing that ensemble methods such as random forests and XGBoost [8], along with artificial neural networks, offer advantages in capturing non-linear relationships and complex interactions. However, a persistent tension exists in sports analytics between predictive accuracy and model interpretability. To ensure that coaching staff and management can trust and act upon model recommendations, researchers have begun to emphasize explainable AI (XAI) tools such as SHAP and LIME [9][10].

NBA Career Longevity and Survival Analysis

Research on NBA career duration has drawn substantially on survival analysis methods. Groothuis and Hill [5] applied duration analysis to study exit discrimination, finding that height and draft position were significant predictors of career longevity, while racial discrimination was not a significant factor in exit patterns. In a subsequent study, Groothuis and Hill [6] examined career duration among foreign-born players, reporting that those without U.S. college experience tended to have shorter careers. Staw and Hoang [11] demonstrated a sunk-cost effect: highly drafted players received more playing time and survived longer in the league, even when their on-court performance did not warrant it. Petersen et al. [12] provided quantitative evidence of the Matthew effect in career longevity, showing that even modest early-career progress plays a critical role in shaping long-term career trajectories.

More recently, Macis [13] applied algorithmic survival data analysis—including survival trees and Random Survival Forests—to NBA player performance measurement, demonstrating the utility of ML methods in handling time-to-event data and capturing non-linear effects. Foutzopoulos et al. [2] directly addressed the prediction of the ten-year full retirement threshold using sophomore-year statistics, finding that Ridge logistic regression achieved an AUC of 0.791 and identifying minutes played and points scored as the most important predictors. The present study extends the work of Foutzopoulos et al. by broadening the feature set to include pre-draft physical measurements, expanding the temporal window to encompass the first three seasons of NBA performance, and introducing more advanced feature selection and model interpretation techniques.

Predictive Value of Draft Combine Data and Early-Career Performance

Whether Draft Combine measurements can reliably predict professional success has been a subject of ongoing debate. Teramoto et al. [14] provided empirical evidence that Combine data—particularly the “length-size” factor reflecting height and wingspan—possessed significant predictive validity for future on-court performance indicators such as defensive box plus/minus and win shares. In contrast, Moxley and Towne [15] argued that, once college performance was controlled for, Draft Combine measurement data contributed little to distinguishing productive from unproductive NBA players, emphasizing instead the stability of basketball skill and game intelligence. Nguyen et al. [16] applied ML and deep learning models to predict NBA player performance and popularity, demonstrating the feasibility of data-driven approaches for multi-faceted player evaluation.

In assessing the contribution of early-career on-court performance to long-term success, rebounding, defense, and efficiency metrics have frequently been identified as relevant predictors. Barnes et al. [17] found that pre-NBA statistics such as rebounds and field goal percentage were significantly associated with career longevity for guards and forwards. Berri [18] proposed the “Wins Produced” model, which emphasized the importance of defensive rebounds, steals, and low turnover rates for team success—a perspective that contrasts with the conventional emphasis on scoring output. Sarlis and Tjortjis [1] further demonstrated through data mining that player position, age, and injury history interact with performance metrics to shape career economics. Through LIME-based feature importance analysis, the present study quantifies the relative contributions of physical measurements and on-court performance indicators in determining long-term career viability.

Methodology

Figure 1 provides an overview of the complete analytical pipeline, spanning data collection through model evaluation and interpretation.

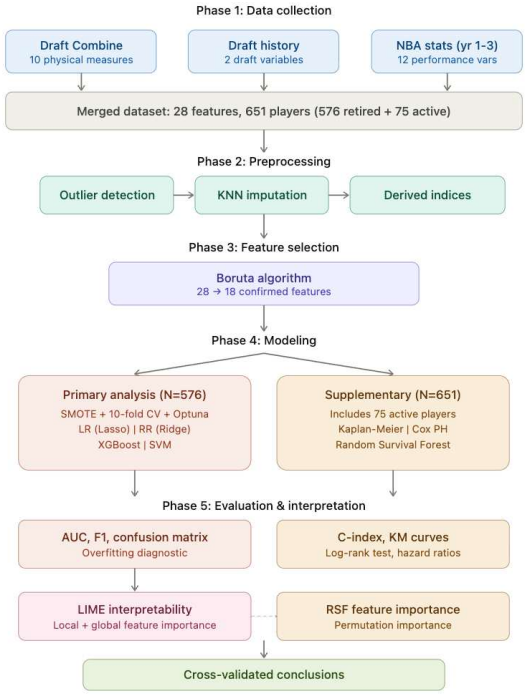


Figure 1. Research framework.

Figure 1. Overview of the proposed analytical framework. The pipeline comprises five phases: (i) multi-source data collection from Draft Combine, draft history, and early-career NBA statistics; (ii) preprocessing via outlier detection, KNN imputation, and derived index computation; (iii) Boruta-based feature selection (28 → 18 features); (iv) dual-track modeling consisting of primary classification on $N = 576$ retired players and supplementary survival analysis on $N = 651$ players including those still active; and (v) evaluation and interpretation using AUC, C-index, LIME, and permutation importance.

Data Collection and Sample

The dataset was assembled using the official NBA statistics API (nba_api) and comprises players who participated in the NBA Draft Combine and were subsequently drafted during the 2000 through 2015 draft classes. The upper bound of 2015 ensures a minimum ten-year

observation window for each player. The final sample consists of 651 players, of whom 576 have retired (event observed) and 75 remain active as of the 2024–25 season (right-censored). Data were obtained from three NBA endpoints: DraftCombinePlayerAnthro and DraftCombineDrillResults for pre-draft physical measurements, DraftHistory for draft information, and LeagueDashPlayerStats for season-level performance statistics.

Variable Definition

A total of 28 predictor variables were assembled from three data sources, as detailed in Table 1.

Table 1. Summary of predictor variables by source.

Source	Variable	Description
Draft Combine (10)	HEIGHT_WO_SHOES	Height without shoes (in.)
	WEIGHT	Body weight (lbs)
	WINGSPAN	Arm wingspan (in.)
	STANDING_REACH	Standing reach height (in.)
	BODY_FAT_PCT	Body fat percentage
	STANDING_VERTICAL_LEAP	Standing vertical leap (in.)
	MAX_VERTICAL_LEAP	Maximum vertical leap (in.)
	LANE_AGILITY_TIME	Lane agility drill time (s)
	THREE_QUARTER_SPRINT	Three-quarter court sprint (s)
	BENCH_PRESS	Bench press repetitions at 185 lbs
Draft Info (2)	OVERALL_PICK	Overall selection number
	ROUND_NUMBER	Draft round (1 or 2)
Derived Indices (4)	BMI	Body mass index
	WINGSPAN_HEIGHT_RATIO	Wingspan-to-height ratio
	POWER_WEIGHT_RATIO	Max vertical leap / weight
	REACH_ADVANTAGE	Standing reach − height
Early-Career NBA (12)	EARLY_AVG_MIN	Mean minutes per game (first 3 seasons)
	EARLY_AVG_PTS	Mean points per game
	EARLY_AVG_REB	Mean rebounds per game
	EARLY_AVG_AST	Mean assists per game

Source	Variable	Description
	EARLY_AVG_STL	Mean steals per game
	EARLY_AVG_BLK	Mean blocks per game
	EARLY_AVG_FG_PCT	Field goal percentage
	EARLY_AVG_FG3_PCT	Three-point field goal percentage
	EARLY_AVG_FT_PCT	Free throw percentage
	EARLY_AVG_TOV	Mean turnovers per game
	EARLY_PEAK_PTS	Peak single-season scoring average
	EARLY_PEAK_MIN	Peak single-season minutes average

The dependent variable, LONGEVITY_10YR, is a binary indicator coded 1 if the player completed ten or more NBA seasons and 0 otherwise. This threshold derives from the NBA CBA, under which ten years of service qualifies a player for the full retirement benefit [2]. Among the 576 retired players, 168 (29.2%) attained this threshold and 408 (70.8%) did not.

Data Preprocessing

Data quality was ensured through a three-stage procedure. First, outliers were identified on the basis of domain knowledge: values falling outside physiologically plausible ranges (e.g., height below 60 or above 95 inches; weight below 140 or above 350 lbs) were set to missing. Second, missing values were imputed according to their rates: variables with missingness at or below 5% were imputed with the median, while those with missingness between 5% and 25% were imputed using k-nearest neighbors ($k = 5$). No variable exceeded 25% missingness after outlier treatment. Third, the four derived indices were recalculated from the cleaned base variables.

Feature Selection

Feature selection was performed using the Boruta algorithm [19], which was chosen both for its statistical rigor in distinguishing genuine predictors from noise through shadow feature testing and for methodological continuity with Foutzopoulos et al. [2], thereby enabling direct comparison of selected feature sets across studies. The algorithm constructs randomized shadow copies of all features, trains a random forest classifier, and iteratively tests whether each original feature is significantly more important than the best-performing shadow feature (max_iter = 100). To mitigate the effect of class imbalance on the selection process, Boruta was applied to SMOTE-resampled training data.

Classification Models

Four classification algorithms were evaluated. The mathematical formulations are provided below.

Logistic Regression with L1 Regularization (LR)

Logistic regression models the probability of the positive class through the logistic function:

$$P(Y = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp\left(-\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right)\right)}$$

where $\mathbf{x} = (x_1, \ldots, x_p)^\top$ denotes the feature vector, β_0 is the intercept, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ are the regression coefficients. With L1 (Lasso) regularization, the coefficients are estimated by minimizing the penalized negative log-likelihood:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)] + \frac{1}{C} \sum_{j=1}^{p} |\beta_j| \right\}$$

where $C > 0$ is the inverse regularization strength optimized via Optuna. The L1 penalty induces sparsity by driving irrelevant coefficients exactly to zero, thereby performing embedded feature selection concurrently with parameter estimation.

Ridge Logistic Regression (RR)

Ridge logistic regression employs an L2 penalty that retains all features while shrinking coefficients toward zero to mitigate multicollinearity:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)] + \frac{1}{C} \sum_{j=1}^{p} \beta_j^2 \right\}$$

This model was included to enable direct methodological comparison with the best-performing algorithm reported by Foutzopoulos et al. [2], who applied Ridge logistic regression to a distinct, single-source feature set comprising sophomore-year statistics.

XGBoost

XGBoost [8] is a gradient-boosted tree ensemble that sequentially constructs K additive regression trees. The predicted output for observation i is:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i), \quad f_k \in \mathcal{F}$$

where $\mathcal{F}$ is the space of regression trees. At each boosting iteration t, a new tree f_t is added to minimize the regularized objective:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)\right) + \Omega(f_t)$$

where l is the logistic loss and $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ penalizes both the number of terminal nodes T and the squared leaf weights w_j. The hyperparameter search space comprised tree depth (2–5), learning rate (0.01–0.15), row subsampling (0.5–0.8), column subsampling (0.4–0.8), L1 penalty α (1–15), L2 penalty λ (2–15), minimum child weight (5–20), and minimum split loss γ (0–5).

Support Vector Machine (SVM)

The SVM [20] seeks a hyperplane that maximizes the margin between classes. For the non-separable case with a radial basis function (RBF) kernel:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{n} \xi_i$$

subject to $y_i(\mathbf{w}^\top \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$, where $\phi(\cdot)$ maps the input to a higher-dimensional space, C controls the margin–misclassification trade-off, and ξ_i are slack variables. The RBF kernel is defined as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \| \mathbf{x}_i - \mathbf{x}_j \|^2\right)$$

Both C and γ were optimized via Optuna. The implementation follows the scikit-learn library [21].

Hyperparameter Optimization

All hyperparameters were optimized using the Optuna framework [22], which implements Bayesian optimization via the Tree-structured Parzen Estimator. Each model underwent 50 optimization trials, with the objective of maximizing 10-fold cross-validated F1 score. To prevent information leakage arising from oversampling prior to cross-validation, SMOTE was integrated

within each fold using the imbalanced-learn Pipeline [23], ensuring that synthetic minority samples were generated exclusively from within-fold training data.

Evaluation Protocol

Model performance was assessed on a stratified 80/20 train–test split (with a fixed random seed for reproducibility), with 10-fold stratified cross-validation conducted on the training partition. Both training-fold and validation-fold F1 scores were recorded to diagnose overfitting (gap = training F1 − validation F1). The primary evaluation metric was the area under the receiver operating characteristic curve (AUC), following the convention of Foutzopoulos et al. [2]. Secondary metrics included F1 score, accuracy, precision, recall, and the confusion matrix.

Model Interpretability

Local Interpretable Model-agnostic Explanations (LIME) [24] were applied to the best-performing model. Three representative instances were selected for case-level analysis: a correctly classified long-career player (≥10 years), a correctly classified short-career player (<10 years), and a misclassified player. Global feature importance was estimated by averaging the absolute LIME weights across 50 randomly sampled test instances. Feature contributions were further aggregated by data source (Early NBA, Draft Combine, Draft Info) to quantify the relative predictive value of each information category.

Supplementary Survival Analysis

To assess the robustness of the classification findings and to accommodate the 75 right-censored (still active) players, supplementary survival analyses were conducted on the full sample ($N = 651$).

Kaplan–Meier Estimation

The Kaplan–Meier estimator provides a non-parametric estimate of the survival function $S(t)$, representing the probability of remaining active in the NBA beyond time t:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where t_i denotes the i-th ordered event time, d_i is the number of retirements at t_i, and n_i is the number of players at risk just prior to t_i. Subgroup comparisons were made by stratifying on

draft position and early-career scoring average, with between-group differences evaluated using the log-rank test.

Cox Proportional Hazards Regression

The Cox model [25] specifies the hazard of career termination at time t for a player with covariate vector $\mathbf{x}$ as:

$$h(t \mid \mathbf{x}) = h_0(t)\exp\left(\sum_{j=1}^{p} \beta_j \, x_j\right)$$

where $h_0(t)$ is the unspecified baseline hazard and $\exp(\beta_j)$ is the hazard ratio (HR) associated with a one-unit increase in x_j. An HR below 1 indicates a protective effect (longer career); an HR above 1 indicates elevated risk of early retirement. L2 penalization (penalizer = 0.1) was applied to stabilize coefficient estimates. Model discrimination was evaluated by the concordance index (C-index).

Random Survival Forest

The Random Survival Forest (RSF) [26] extends random forests to right-censored data. It constructs $B = 500$ trees, each grown on a bootstrap sample. At each node, the log-rank splitting rule identifies the covariate and split point that maximize the survival difference between child nodes. The cumulative hazard function for terminal node W_k is estimated by the Nelson–Aalen estimator:

$$\widehat{H}_k(t) = \sum_{t_i \le t} \frac{d_i^{(k)}}{n_i^{(k)}}$$

where $d_i^{(k)}$ and $n_i^{(k)}$ denote the number of events and at-risk individuals at time t_i within node W_k. The ensemble prediction is obtained by averaging cumulative hazard estimates across all trees. Variable importance was assessed via permutation importance on the held-out test set (20 repetitions). The C-index was reported on the same test partition.

Results

Feature Selection

Boruta confirmed 18 of 28 features as statistically relevant (Figure 2). All early-career NBA performance variables except three-point field goal percentage were selected. Six physical measurement variables were confirmed: wingspan, maximum vertical leap, lane agility time, three-quarter court sprint, BMI, and wingspan-to-height ratio. Overall draft pick was the sole selected draft variable. Five features were classified as tentative (reach advantage, bench press, standing vertical leap, body fat percentage, and height), and five were rejected (weight, standing reach, round number, power-to-weight ratio, and three-point field goal percentage). Results are summarized in Table 2.

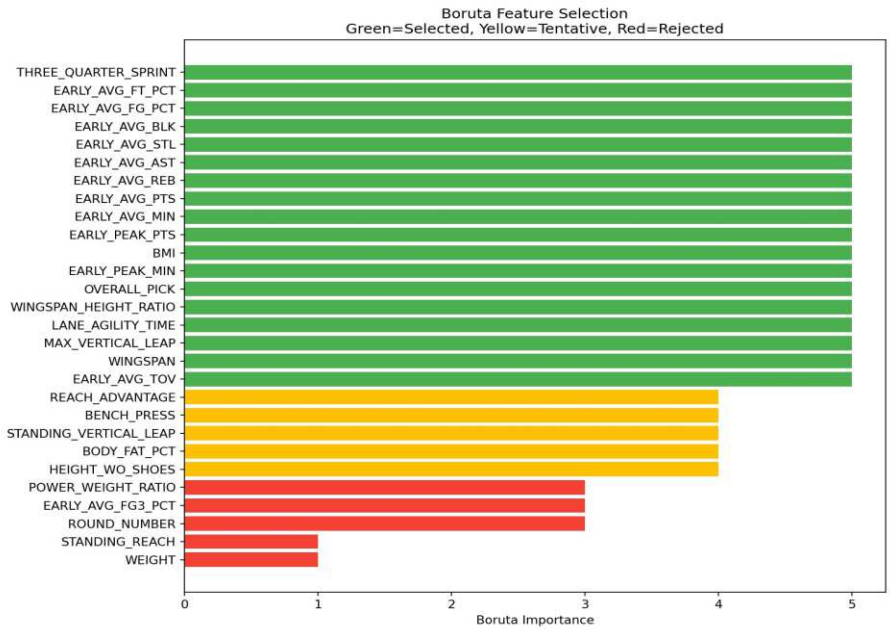


Figure 2. Boruta feature selection results.

Figure 2. Boruta feature selection results. Green bars denote confirmed features, yellow bars tentative features, and red bars rejected features.

Table 2. Boruta feature selection results.

Decision	Features	Count
Confirmed	EARLY_AVG_MIN, EARLY_AVG_PTS, EARLY_AVG_REB, EARLY_AVG_AST, EARLY_AVG_STL, EARLY_AVG_BLK, EARLY_AVG_FG_PCT, EARLY_AVG_FT_PCT, EARLY_AVG_TOV, EARLY_PEAK_PTS, EARLY_PEAK_MIN, WINGSPAN,	18

Decision	Features	Count
	MAX_VERTICAL_LEAP, LANE_AGILITY_TIME, THREE_QUARTER_SPRINT, BMI, WINGSPAN_HEIGHT_RATIO, OVERALL_PICK	
Tentative	REACH_ADVANTAGE, BENCH_PRESS, STANDING_VERTICAL_LEAP, BODY_FAT_PCT, HEIGHT_WO_SHOES	5
Rejected	WEIGHT, STANDING_REACH, ROUND_NUMBER, POWER_WEIGHT_RATIO, EARLY_AVG_FG3_PCT	5

Classification Performance

Table 3 reports the performance of all four models. All achieved test AUC values exceeding 0.80, indicating good discriminative ability. L1 logistic regression (LR) attained the highest test AUC of 0.840, followed by Ridge regression (0.829), XGBoost (0.816), and SVM (0.804). Overfitting was minimal across all models: the largest F1 gap was 0.040 (SVM), well below the conventional threshold of 0.10. LR exhibited a negative gap of −0.011, indicating that validation performance marginally exceeded training performance (Figure 3).

Table 3. Classification performance across all models.

Model	CV F1 (train)	CV F1 (val)	Gap	CV AUC	Test AUC	Test F1	Test Acc	Test Prec	Test Rec
LR (Lasso)	0.643	0.654	−0.011	0.819	**0.840**	0.608	0.733	0.533	0.706
RR (Ridge)	0.667	0.647	+0.020	0.818	0.829	0.571	0.690	0.480	0.706
XGBoost	0.670	0.652	+0.018	0.808	0.816	0.634	0.741	0.542	0.765
SVM	0.692	0.652	+0.040	0.813	0.804	0.612	0.716	0.510	0.765

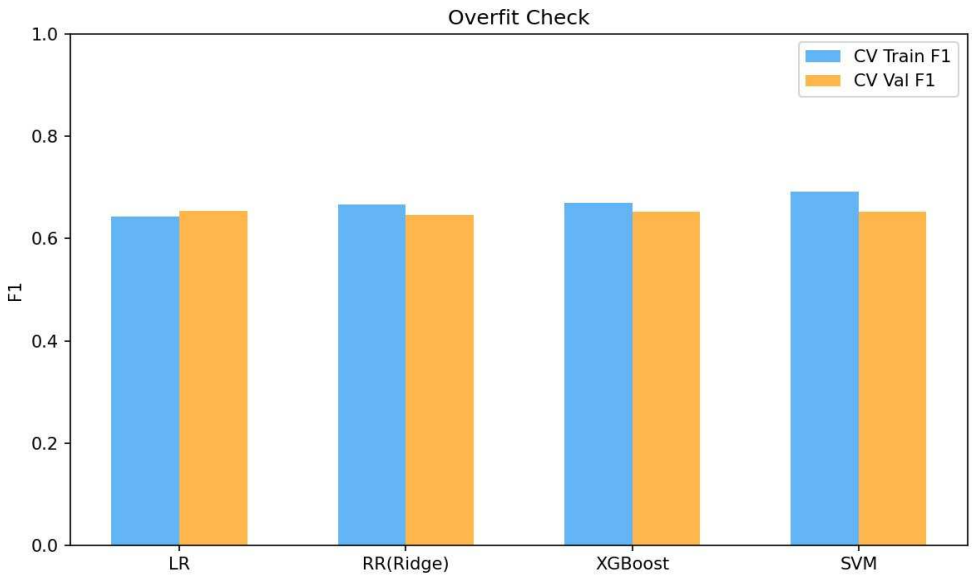


Figure 3. Overfitting diagnostic.

Figure 3. Overfitting diagnostic. Blue bars represent CV training F1; orange bars represent CV validation F1. The negligible gap across all models indicates well-controlled generalization.

Figure 4 presents the ROC curves. LR (solid blue) consistently achieves the highest sensitivity at most specificity levels.

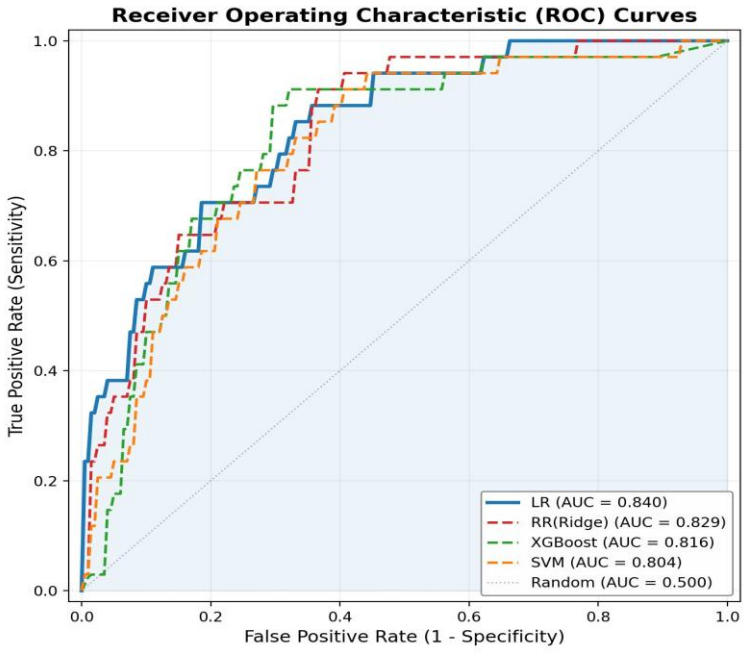


Figure 4. ROC curves.

Figure 4. Receiver operating characteristic curves. LR = L1 logistic regression (AUC = 0.840); RR = Ridge regression (AUC = 0.829); XGBoost (AUC = 0.816); SVM (AUC = 0.804). The dotted diagonal represents chance-level performance.

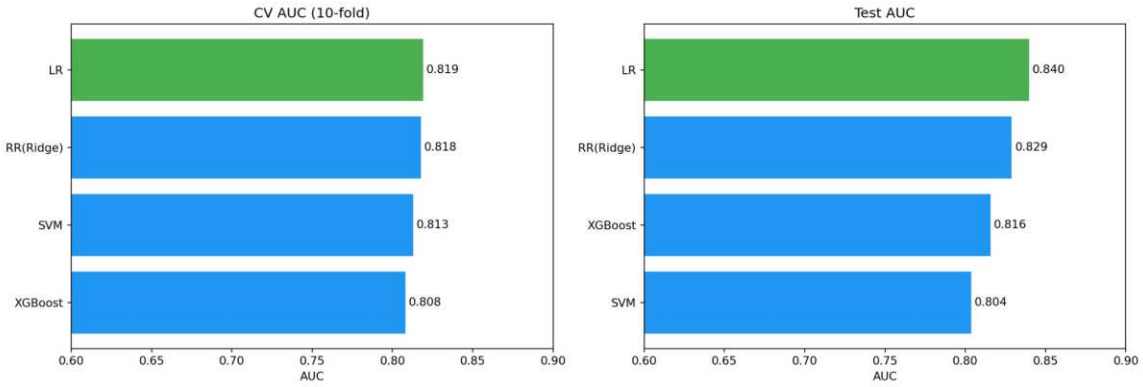


Figure 5. AUC comparison.

Figure 5. AUC comparison. Left panel: 10-fold cross-validated AUC. Right panel: held-out test set AUC. Green bars indicate the best-performing model (LR).

On the held-out test set ($N = 116$), LR achieved an accuracy of 73.3%, a sensitivity of 70.6%, and a specificity of 74.4%. The confusion matrix (Figure 6) shows 61 true negatives, 24 true positives, 21 false positives, and 10 false negatives.

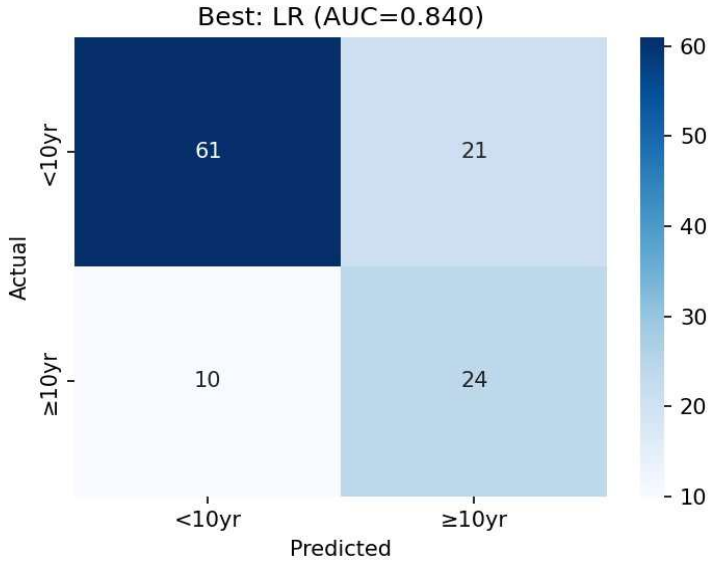


Figure 6. Confusion matrix.

Figure 6. Confusion matrix for the best-performing model (LR, AUC = 0.840). The model correctly identified 24 of 34 long-career players and 61 of 82 short-career players.

Model Interpretability

LIME analysis of the LR model revealed that early-career NBA performance variables accounted for 85.2% of the total predictive contribution, Draft Combine physical measurements for 12.9%, and draft information for 1.9% (Figure 7; Table 4).

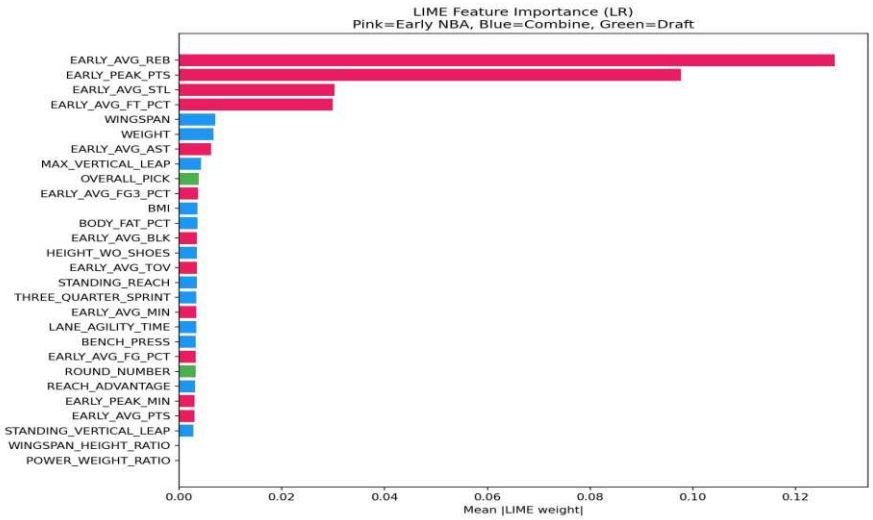


Figure 7. Global LIME feature importance.

Figure 7. Global LIME feature importance (LR model), aggregated over 50 test instances. Pink = Early NBA performance; blue = Draft Combine measurements; green = Draft information.

Table 4. LIME feature importance (top 10) and contribution by source.

Rank	Feature	LIME Weight	Source
1	EARLY_AVG_REB	0.1276	Early NBA
2	EARLY_PEAK_PTS	0.0977	Early NBA
3	EARLY_AVG_STL	0.0302	Early NBA
4	EARLY_AVG_FT_PCT	0.0299	Early NBA
5	WINGSPAN	0.0070	Combine
6	WEIGHT	0.0067	Combine
7	EARLY_AVG_AST	0.0062	Early NBA
8	MAX_VERTICAL_LEAP	0.0043	Combine

Rank	Feature	LIME Weight	Source
9	OVERALL_PICK	0.0038	Draft
10	EARLY_AVG_FG3_PCT	0.0037	Early NBA

Case-level LIME analysis (Figure 8) provided further insight. For a correctly classified long-career player (panel a), high values of early-career rebounds and peak scoring were the dominant positive contributors. For a correctly classified short-career player (panel b), low peak scoring and low rebounds drove the negative prediction. In a misclassified case (panel c), the model was led astray by a high peak scoring value that did not translate into sustained career longevity, likely reflecting unobserved factors such as injury.

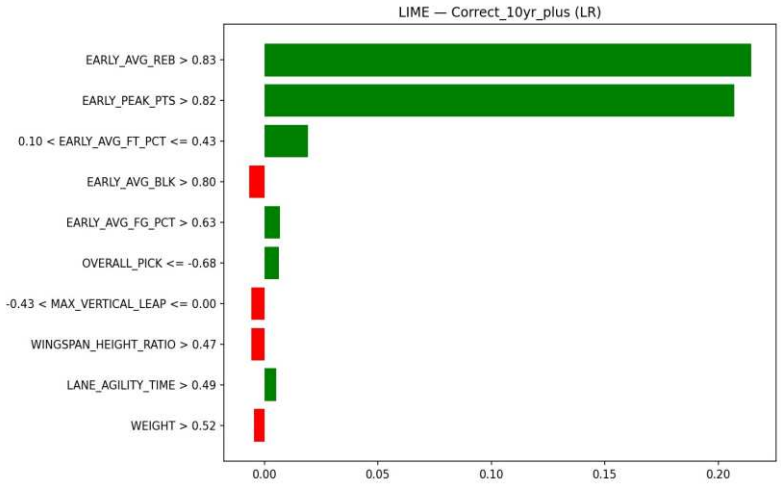


Figure 8a. LIME: correct long-career prediction.

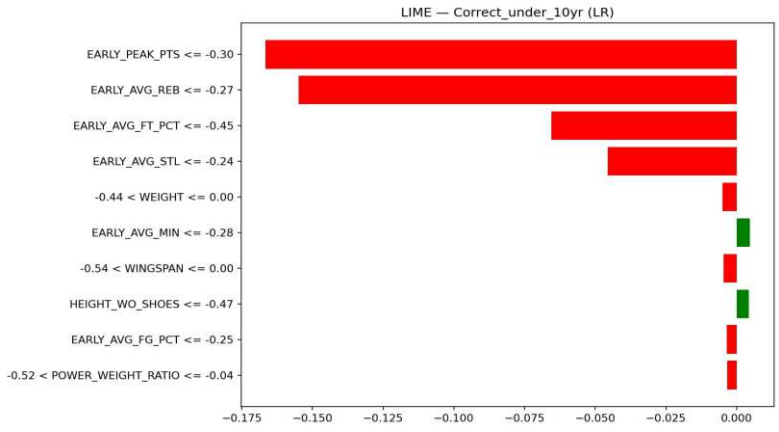


Figure 8b. LIME: correct short-career prediction.

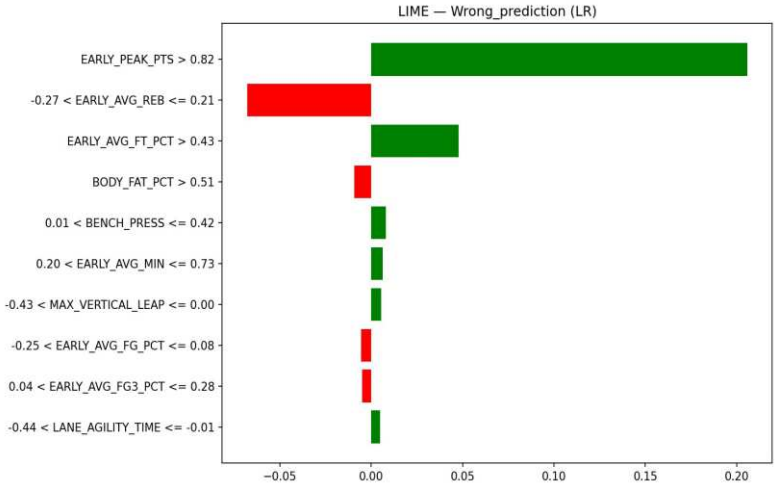


Figure 8c. LIME: incorrect prediction.

Figure 8. LIME explanations for three individual cases. (a) Correct prediction of a ≥ten-year career. (b) Correct prediction of a <ten-year career. (c) Misclassification: high peak scoring misled the model.

Supplementary Survival Analysis

Kaplan–Meier estimation yielded a ten-year survival probability of 0.310 (Figure 9a), consistent with the 29.2% prevalence in the retired subsample. Both stratification variables produced highly significant separation: high draft picks (≤28) versus low picks (>28; $p < 0.001$; Figure 9b), and high early-career scoring (≥5.3 ppg) versus low scoring (<5.3; $p < 0.001$; Figure 9c).

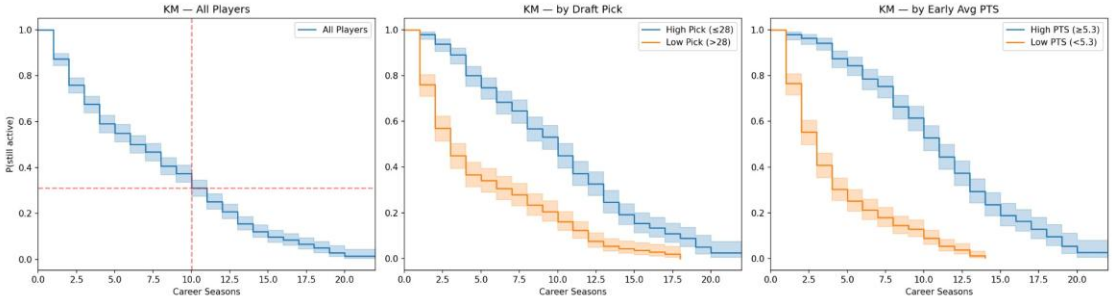


Figure 9. Kaplan-Meier survival curves.

Figure 9. Kaplan–Meier survival curves. (a) Overall survival function; red dashed lines mark the ten-year threshold ($P = 0.310$). (b) Stratification by draft pick. (c) Stratification by early-career scoring. Both comparisons significant at $p < 0.001$ (log-rank test). Shaded bands represent 95% confidence intervals.

Cox proportional hazards regression yielded a C-index of 0.811. Seven features attained statistical significance at $p < 0.05$ (Table 5; Figure 10). Peak minutes exhibited the strongest protective association (HR = 0.615, $p = 0.001$), corresponding to a 38.5% reduction in the hazard of career termination per standard deviation increase. Turnovers were the sole risk factor (HR = 1.230, $p = 0.040$).

Table 5. Cox proportional hazards model: significant features ($p < 0.05$).

Feature	Coefficient	Hazard Ratio	p-value	Direction
EARLY_AVG_FG_PCT	−0.143	0.866	<0.001	Protective
EARLY_PEAK_MIN	−0.486	0.615	0.001	Protective
EARLY_AVG_FT_PCT	−0.116	0.890	0.006	Protective
EARLY_PEAK_PTS	−0.334	0.716	0.009	Protective
EARLY_AVG_FG3_PCT	−0.260	0.771	0.034	Protective
EARLY_AVG_TOV	+0.207	1.230	0.040	Risk
EARLY_AVG_REB	−0.171	0.843	0.042	Protective

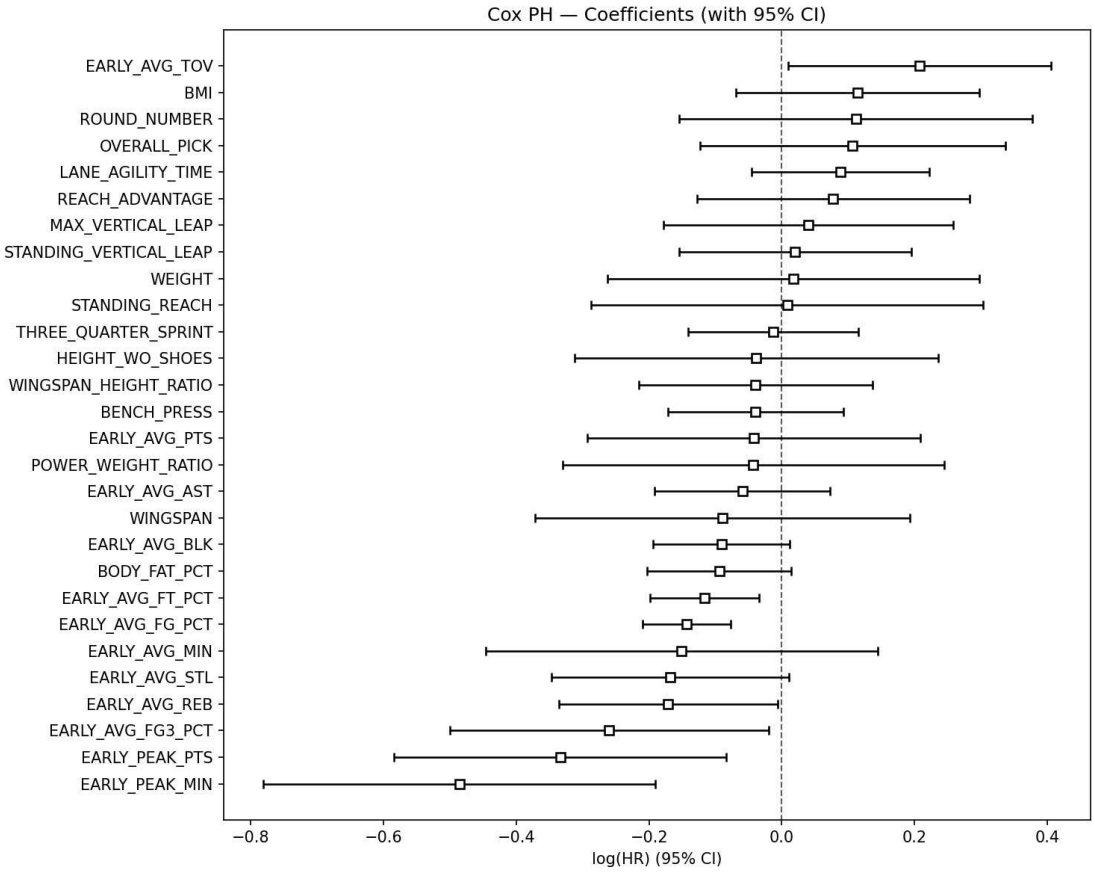


Figure 10. Cox PH coefficients.

Figure 10. Cox proportional hazards model coefficients with 95% confidence intervals. Coefficients to the left of the dashed line at zero indicate protective effects (longer careers).

The Random Survival Forest yielded a C-index of 0.768. Permutation importance identified blocks, field goal percentage, minutes, peak minutes, and rebounds as the five most important features (Figure 11). Aggregated by source, early-career NBA variables contributed 71.7% of the total positive importance, physical measurements 24.2%, and draft information 4.1%.

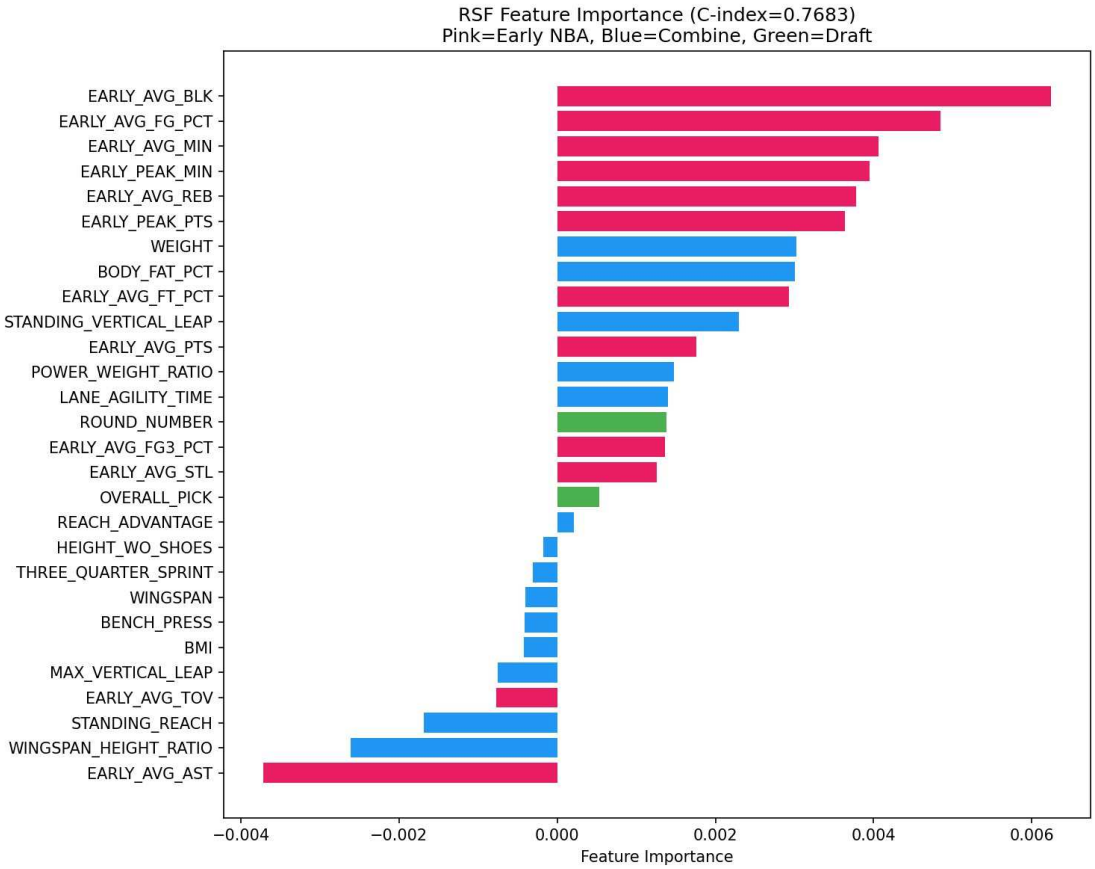


Figure 11. RSF feature importance.

Figure 11. Random Survival Forest permutation importance, colored by data source. Pink = Early NBA; blue = Draft Combine; green = Draft Info. C-index = 0.768.

Discussion

Model Performance and Comparison with Prior Work

The L1-regularized logistic regression model achieved a test AUC of 0.840, compared with the AUC of 0.791 reported by Foutzopoulos et al. [2] using Ridge logistic regression on sophomore-year statistics alone. Although both studies address the same institutional threshold, the present work differs in several important respects: it draws on three data sources rather than one (Draft Combine measurements, draft information, and early-career NBA statistics vs. sophomore-year statistics only), covers a larger and more recent sample (576 players from the 2000–2015 draft classes vs. 322 from 1999–2006), and employs a broader feature set (28 vs. 18 variables). The observed performance gain likely reflects these methodological extensions—particularly the multi-source data fusion and the use of three-season averages, which attenuate single-year noise. Direct numerical comparison should nonetheless be interpreted with caution given these design differences.

Consistent with Foutzopoulos et al. [2], the present study found that regularized linear models outperformed more complex non-linear alternatives on the test set. This pattern has been observed in other sports prediction contexts: Ievoli et al. [27] and Cao [28] both reported that logistic regression matched or exceeded the performance of random forests and neural networks in football outcome prediction. The consistency of this finding across studies and sports suggests that the relationship between early-career indicators and long-term career outcomes is predominantly linear, and that the added flexibility of non-linear models yields limited benefit relative to the elevated risk of overfitting, particularly at moderate sample sizes.

In the supplementary survival analysis, Cox regression (C-index = 0.811) also outperformed the Random Survival Forest (C-index = 0.768). This result parallels that of Macis [13], who, although addressing a different research question (single-season scoring thresholds rather than career longevity), similarly observed lower prediction error for Lasso Cox regression (13.2%) than for RSF (15.2%). The convergence of findings across classification and survival frameworks reinforces the conclusion that regularized linear models are well suited to career longevity prediction in professional basketball.

Feature Importance and Practical Implications

LIME analysis indicated that early-career on-court performance accounted for 85.2% of the total predictive contribution, with physical measurements contributing a secondary but non-negligible

12.9%. This finding refines the conclusions of Moxley and Towne [15], who reported that physical attributes contributed little beyond college performance in distinguishing productive and unproductive NBA players. While our results corroborate the secondary role of Draft Combine data, they also demonstrate that specific attributes—notably wingspan, vertical leap, and agility—retain independent predictive value even after conditioning on three seasons of professional performance. This is broadly consistent with Teramoto et al. [14], who reported long-term predictive validity for certain Combine measurements.

Early-career average rebounds emerged as the single most important predictor. This observation aligns with a longstanding finding in the NBA career literature: defensive contributions, particularly rebounding, positively influence career longevity [11][5]. Foutzopoulos et al. [2] identified minutes played and points scored as the most important features; our results confirm the role of scoring (peak points ranked second) while elevating rebounds to the foremost position. Rebounding captures not only physical capacity but also positional effort and team contribution beyond scoring—what Berri [18] termed “production of wins.”

The Cox model identified turnovers as the sole significant risk factor (HR = 1.230). Although Foutzopoulos et al. [2] observed a positive association between turnovers and career duration (attributing it to the correlation between ball handling and playing time), the Cox model in the present study controls simultaneously for minutes and scoring, thereby isolating the net negative effect of ball mishandling. This indicates that turnovers, independent of usage, constitute a genuine liability for career sustainability.

Boruta rejected three-point percentage, and the Cox model assigned it only marginal significance ($p = 0.034$). While this may appear counterintuitive given the modern NBA’s emphasis on perimeter shooting, it suggests that career longevity depends more on versatility and fundamental skills—rebounding, defense, shooting efficiency—than on a single specialized ability. The RSF analysis, however, assigned greater weight to physical measurements (24.2% vs. 12.9% from LIME), with blocks ranking first in permutation importance. This discrepancy, consistent with findings from Macis [13], suggests that tree-based methods may detect non-linear interactions among physical attributes—such as threshold effects involving wingspan and shot-blocking—that linear models cannot represent.

Methodological Contributions

This study offers four methodological contributions to the sports data mining literature. First, the restriction of SMOTE oversampling to within cross-validation folds prevents the artificial inflation of performance estimates—a pitfall often overlooked in earlier ML-based sports prediction studies [23]. Second, the use of Bayesian hyperparameter optimization via Optuna [22] provides a principled and reproducible basis for model comparison. Third, the combination of Boruta feature selection and LIME interpretability yields complementary views: Boruta identifies globally relevant features via statistical testing, while LIME reveals the feature-level mechanisms underlying individual predictions [19][24]. Fourth, the supplementary survival analysis—conducted using Cox regression [25] and RSF [26]—serves as an independent validation of the classification results and appropriately handles right-censored data from active players.

Limitations

Several limitations warrant acknowledgment. First, the sample is restricted to players who participated in the NBA Draft Combine and were subsequently drafted; international players entering through alternative pathways are excluded. As Groothuis and Hill [6] noted, the career trajectories of foreign-born players may be governed by distinct factors. Second, the model does not incorporate injury data. Substantial evidence indicates that major injuries—including ACL tears [29], Achilles tendon ruptures [30], and rookie-season ailments [31]—substantially curtail career duration. Third, the ten-year threshold, while institutionally significant, represents a single dichotomization of a continuous outcome; alternative thresholds may yield different findings. Fourth, the NBA is a continuously evolving system; changes in playing style, rules, and economic structure may limit the temporal generalizability of the model. Fifth, the sample size ($N = 576$ for the primary analysis), while larger than those of comparable studies, may still constrain the capacity of complex models to generalize effectively.

Conclusion

This study developed a multi-algorithm machine learning framework for predicting whether NBA players achieve a career of ten or more seasons, using pre-draft physical measurements, draft information, and early-career performance data. L1-regularized logistic regression emerged as the optimal model (AUC = 0.840), outperforming Ridge regression (0.829), XGBoost (0.816), and SVM (0.804), with minimal overfitting observed across all models. These results both

replicate and extend the findings of Foutzopoulos et al. [2], confirming the suitability of regularized linear models for this prediction task and demonstrating the value of multi-source data integration.

Feature analysis consistently identified early-career rebounds, peak scoring, steals, and shooting efficiency as the strongest predictors of career longevity, with physical attributes contributing a secondary but statistically meaningful role (12.9% by LIME; 24.2% by RSF). Supplementary survival analysis via Cox regression (C-index = 0.811) and RSF (C-index = 0.768) corroborated these findings while properly accommodating right-censored observations from 75 active players.

For NBA franchises, the results suggest that early-career defensive contributions (rebounds, blocks, steals), shooting efficiency, and low turnover rates are more indicative of long-term viability than raw scoring output or pre-draft physical measurements alone. For the sports analytics community, the consistent finding that parsimonious, interpretable models match or surpass complex alternatives reinforces the principle that model complexity should be justified by demonstrated performance gains rather than adopted by default.

Future work should incorporate injury histories, psychological assessments, and team-level contextual variables. Temporal validation using rolling windows, as recommended by Foutzopoulos et al. [2], would further evaluate model robustness in the face of the NBA's ongoing evolution. The stronger contribution of physical measurements in the RSF analysis (24.2% vs. 12.9%) also warrants deeper investigation into non-linear interactions between pre-draft physical attributes and career outcomes.

Data Availability Statement

The data supporting the findings of this study were obtained from the publicly available NBA statistics API (https://github.com/swar/nba_api). The analysis code and processed datasets are available from the corresponding author upon reasonable request.

Author Contributions Statement

R.L. conceived and designed the study, collected and processed the data, performed the machine learning analysis, and drafted the manuscript. Y.G. contributed to data preprocessing and validation of results. Y.Z. assisted with the survival analysis implementation. L.W.

supervised the research, provided critical revisions, and approved the final version. All authors reviewed the manuscript.

Competing Interests

The authors declare no competing interests.

Funding

This research received no external funding.

Acknowledgements

The authors would like to thank the NBA for making statistical data publicly available through its official API.

Ethics Statement

This study exclusively used publicly available, anonymized statistical data obtained from the NBA API. No human participants were recruited, no personally identifiable information was collected, and no experimental interventions were performed. Ethical approval was therefore not required for this study.

Additional Information

Correspondence and requests for materials should be addressed to L.W. (13153038192@163.com).

References

[1] Sarlis, V., & Tjortjis, C. (2024). Sports analytics: Data mining to uncover NBA player position, age, and injury impact on performance and economics. *Information*, 15(4), 242. https://doi.org/10.3390/info15040242

[2] Foutzopoulos, G., Pandis, N., & Tsagris, M. (2025). Predicting full retirement attainment of NBA players. *International Journal of Data Science and Analytics*, 20, 6121–6134. https://doi.org/10.1007/s41060-025-00821-z

[3] Sarlis, V., & Tjortjis, C. (2020). Sports analytics—Evaluation of basketball players and team performance. *Information Systems*, 93, 101562. https://doi.org/10.1016/j.is.2020.101562

[4] Bunker, R., & Susnjak, T. (2022). The application of machine learning techniques for predicting match results in team sport: A review. *Journal of Artificial Intelligence Research*, 73, 1285–1322. https://doi.org/10.1613/jair.1.13509

[5] Groothuis, P. A., & Hill, J. R. (2004). Exit discrimination in the NBA: A duration analysis of career longevity. *Economic Inquiry*, 42(2), 341–349. https://doi.org/10.1093/ei/cbh065

[6] Groothuis, P. A., & Hill, J. R. (2018). Career duration in the NBA: Do foreign players exit early? *Journal of Sports Economics*, 19(6), 873–883. https://doi.org/10.1177/1527002517690789

[7] Foteinakis, P. F., Kokkotis, C., Karamousalidis, G., et al. (2025). From data to decisions: Using explainable machine learning to predict EuroLeague basketball outcomes. *Applied Sciences*, 15(23), 12401. https://doi.org/10.3390/app152312401

[8] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

[9] Odet, A., et al. (2025). A machine learning and explainability-driven methodology for sports analytics. *Expert Systems with Applications*, 277, 125000. https://doi.org/10.1016/j.eswa.2025.125000

[10] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.

[11] Staw, B. M., & Hoang, H. (1995). Sunk costs in the NBA: Why draft order affects playing time and survival in professional basketball. *Administrative Science Quarterly*, 40(3), 474–494. https://doi.org/10.2307/2393794

[12] Petersen, A. M., Jung, W.-S., Yang, J.-S., & 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. https://doi.org/10.1073/pnas.1016733108

[13] Macis, A. (2025). Basketball players performance measurement with algorithmic survival data analysis. *AStA Advances in Statistical Analysis*, 109, 529–555. https://doi.org/10.1007/s10182-025-00533-6

[14] Teramoto, M., Cross, C. L., Rieger, R. H., Maak, T. G., & Willick, S. E. (2018). Predictive validity of National Basketball Association draft combine on future performance. *Journal of Strength and Conditioning Research*, 32(2), 396–408. https://doi.org/10.1519/JSC.0000000000001798

[15] Moxley, J. H., & Towne, T. J. (2015). Predicting success in the National Basketball Association: Stability & potential. *Psychology of Sport and Exercise*, 16, 128–136. https://doi.org/10.1016/j.psychsport.2014.07.003

[16] Nguyen, N. H., Nguyen, D. T. A., & Ma, B. (2022). The application of machine learning and deep learning in sport: Predicting NBA players' performance and popularity. *Journal of Information and Telecommunication*, 6(2), 217–235. https://doi.org/10.1080/24751839.2021.1977066

[17] Barnes, J. C., Abrams, W., & Clement, A. (2008). Relationship of selected pre–NBA career variables to NBA players' career longevity. *The Sport Journal*, 11(2).

[18] Berri, D. J. (1999). Who is 'most valuable'? Measuring the player's production of wins in the National Basketball Association. *Managerial and Decision Economics*, 20(8), 411–427. https://doi.org/10.1002/1099-1468(199912)20:8<411::AID-MDE957>3.0.CO;2-G

[19] Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11), 1–13. https://doi.org/10.18637/jss.v036.i11

[20] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297. https://doi.org/10.1007/BF00994018

[21] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[22] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2623–2631). https://doi.org/10.1145/3292500.3330701

[23] Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17), 1–5.

[24] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

[25] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2), 187–220. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

[26] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *Annals of Applied Statistics*, 2(3), 841–860. https://doi.org/10.1214/08-AOAS169

[27] Ievoli, R., Palazzo, L., & Ragozini, G. (2021). On the use of passing network indicators to predict football outcomes. *Knowledge-Based Systems*, 222, 106997. https://doi.org/10.1016/j.knosys.2021.106997

[28] Cao, S. (2024). Passing path predicts shooting outcome in football. *Scientific Reports*, 14(1), 9572. https://doi.org/10.1038/s41598-024-60193-1

[29] Kester, B. S., Behery, O. A., Minhas, S. V., & 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. https://doi.org/10.1007/s00167-016-4060-y

[30] Khalil, L. S., et al. (2020). Effect of Achilles tendon rupture on player performance and longevity in National Basketball Association players. *Orthopaedic Journal of Sports Medicine*, 8(11). https://doi.org/10.1177/2325967120966041

[31] Martin, C. L., et al. (2021). Characterization of rookie season injury and illness and career longevity among National Basketball Association players. *JAMA Network Open*, 4(10), e2128670. https://doi.org/10.1001/jamanetworkopen.2021.28670

[32] Miguel, C. G., et al. (2019). Modelling the relationship between NBA draft and the career longevity of players using generalized additive models. *Revista de Psicología del Deporte*, 28(3), 65–70.

[33] Sarlis, V., et al. (2025). Evaluating the effectiveness of machine learning models for performance forecasting in basketball: A comparative study. *Knowledge and Information Systems*, 66, 4395–4435. https://doi.org/10.1007/s10115-024-02092-9