Modeling Career Longevity in NBA Players: An Analysis of Draft Position and Performance Metrics*

For players drafted between 1990 and 2010, which career performance metrics are most strongly associated with career longevity, controlling for draft position?

Arye Santosh

December 10, 2025

This paper uses simple linear regression to test whether driving or putting is a stronger linear predictor of scoring average on the PGA Tour. Using aggregated 2022 player data, scoring average is modeled as a function of average strokes gained off-the-tee and average strokes gained putting. The analysis reveals that while both predictors have a statistically significant negative relationship with scoring average, the model with putting performance has a higher R-squared than driving performance and explains more of the variability. From the data and analysis, it can be concluded that putting performance is a stronger linear predictor of average score for professional golfers.

Introduction

The National Basketball Association (NBA) is a multi-billion dollar industry where talent acquisition and player development represent significant investments. Each year, teams spend resources on scouting, drafting, and training young players, with the hope that their investments will pay off long term. Career longevity - the number of years a player remains in the league - is a critical measure of player value, impacting team continuity, salary management, and organizational planning. Understanding which factors are most strongly associated with longer careers can potentially provide insights to not only what teams can further investments in, but also what players should focus on to maximize the span of their career.

Previous research on NBA player performance examines many metrics of career success, such as All-Star selections, statistical production peaks, and of course, salary. Studies have shown

*Project repository available at: https://github.com/arye69/MATH261A-p2.

draft position serves as a strong predictor of initial playing opportunities and early career success. However, the relationship between draft position, career performance metrics, and career longevity remains obscure. Simple bivariate comparisons or using multiple correlated performance variables without addressing multicollinearity is common in existing models. This gap limits the understanding of which factors independently affect career longevity, while controlling for draft position.

This paper addresses said gap by comparing two modeling approaches: a full multiple linear regression model and a LASSO (Least Absolute Shrinkage and Selection Operator) regularized model, both applied to NBA players drafted between 1990 and 2010. We use career performance statistics to conduct an explanatory analysis of factors associated with longevity, acknowledging that these statistics are not available at draft time for true prediction. The LASSO model automatically selects the most important variables while shrinking less explanatory coefficients to 0, resulting in a more parsimonious model. Both models are evaluated using holdout validation on temporally separated test data to assess their comparative performance and generalizability.

The remainder of the paper is as follows. The `Data` section goes over the NBA draft dataset in detail including its source, relevant variables, and an explanation of pertinent metrics. The `Methods` section defines the multiple linear regression method used and describes the process of evaluating model fit and assumption checks. Finally, the `Results` section presents and assesses the findings of the fitted model comparisons between a full model (with all predictors) and a LASSO selected model.

Data

The data for this analysis was obtained from the SCORE Sports Data Repository which sourced its data from a Kaggle dictionary, the author of which sourced the data from the National Collegiate Athletic Association (Vivian Johnson, NCAA, and Ben Wieland 2024).

The observational unit in the raw dataset (`nba_draft`) is a basketball player. The data was collected in 2021 with every player in the dataset having all their stats current to that year. Each observation (and player) is unique, so out of the 1621 observations, each one represents an individual player and their stats. No player is repeated in the dataset.

Relevant variables

The following variables are relevant:

- `years_played`: The response variable denoting the number of years a player has been in the league for, since they were drafted

- `mins_per_game`: average number of minutes played per game by a player during their NBA career.
- `pts_per_game`: average number of points scored per game by a player during their NBA career.
- `rebounds_per_game`: average number of rebounds completed per game by a player during their NBA career.
- `assists_per_game`: average number of assists completed per game by a player during their NBA career.
- `fg_percent`: percentage of field goals made by a player during their NBA career.
- `three_pt_percent`: percentage of three point shots made by a player during their NBA career.
- `ft_percent`: percentage of free throws made by a player during their NBA career.
- `pick_overall`: The overall pick number a player was selected during the NBA draft for their year. Note this is labelled as `draft_pick` in the data dictionary, but called `pick_overall` in the actual dataset.

Data cleaning

The data cleaning process from the raw dataset to the used one was fairly simple. First, all players with a draft year after 2010 were filtered out. There's a high likelihood that players drafted after that would still be part of the league, which would defeat the purpose of predicting career longevity. The `pick_overall` variable was modified to be a categorical one, labelled `draft_cat`. `draft_cat` was coded to label players with draft numbers 1-14 as "Lottery" (Lottery picks), players drafted 15-30 as "Mid-first", players drafted 31-45 as "Early-second", and those drafted 46+ as "Late-second". The use of a categorical variable allowed for group generalizations by using draft pick as a predictor. Dplyr was used to clean this dataset (Wickham et al. 2023).

Limitations

A potential limitation of the data is measurement error. As with anything sports related, no measurements are perfect, but they're as good as they'll get since they're from official NCAA data. Era effects are also lost. The NBA has changed a lot from 1990 to 2010 and those changes in playing style, rule modification, league expansion and other things of the sort likely affect the data. Another thing that's absent from the data is contextual variables, such as position, physical attributes, possible injuries, and anything else that would affect the metrics used to build the models in this paper. Finally, the practice of using career metrics to predict career

longevity is definitely circular, but the models are mean for direct comparison as opposed to true prediction.

Methods

Both models followed a multiple linear regression framework, using a different set of predictors to model career longevity. The general model form is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_n X_{in} + \varepsilon$$

Where

- Y_i represents career longevity (in years) for player i
- β_0 is the intercept term, denoting expected career longevity when all other predictors are 0
- $\beta_1, \beta_2, \ldots, \beta_n$ are the regression coefficients for each predictor
- $X_{i1}, X_{i2}, \ldots, X_{in}$ are the predictor variables for player i
- ε is the random error term, denoting variance not captured by the model

The full model will be structured as follows:

$$\begin{aligned}\text{years\_played}_i = \beta_0 &+ \beta_1 \text{mins\_per\_game}_i + \beta_2 \text{pts\_per\_game}_i \\ &+ \beta_3 \text{rebounds\_per\_game}_i + \beta_4 \text{assists\_per\_game}_i \\ &+ \beta_5 \text{fg\_percent}_i + \beta_6 \text{three\_pt\_percent}_i \\ &+ \beta_7 \text{ft\_percent}_i + \sum_{k=1}^{3} \gamma_k \text{draft\_category}_{ik} + \epsilon_i\end{aligned}$$

where γ_k coefficients represent the effects of the categorical `draft_cat` variable (with the "Lottery" level as reference). The LASSO regularized model's objective function is:

$$\min_{\beta_0, \beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

where n is the number of observations, p is the number of predictors, and $\lambda \geq 0$ is the tuning parameter that controls the strength of regularization. The optimal λ was selected using 10-fold cross-validation, minimizing the mean square error (MSE). The one standard-error lambda was utilized in this report as opposed to the minimum lambda, since the most parsimonious LASSO model was desired.

Variable selection and justification

The candidate predictors consisted of seven performance metrics and one categorical variable. The continuous predictors were `mins_per_game`, `pts_per_game`, `rebounds_per_game`, `assists_per_game`, `fg_percent`, `three_pt_percent`, and `ft_percent`. The categorical variable was `draft_cat`, added to the original dataset in processing. It has four levels, further described in the **Data** section.

The data was partitioned using a random 80-20 training-test split to create a training set with 829 observations and a test set with 197 observations. The split was performed with random seed (seed = 69) for reproducibility.

Model Validation

The models were evaluated using multiple criteria. The in-sample fit metrics used were:

- R^2 (coefficient of determination) and adjusted R^2
- Akaike Information Criterion (AIC): AIC $= 2k - 2\ln(\hat{L})$, where k =number of parameters
- Bayesian Information Criterion (BIC): BIC $= k\ln(n) - 2\ln(\hat{L})$

The out-of-sample fit metrics used were:

- Root Mean Square Error (RMSE): $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
- Mean Absolute Error (MAE): $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- Bias: $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$

Software

All statistical analysis was done in R (R Core Team 2024). The here package was used to import and load files (Müller 2025). The lm() function in base R was utilized to fit the linear regression model and further diagnostic checks were generated from the same function output. The cv.glmnet() function from the glmnet package (J, T, and R 2010) was also used for LASSO regression. Data manipulation and cleaning was performed using the Tidyverse package (Wickham et al. 2019). All plots and visualizations were created using ggplot2 (Wickham 2016) unless otherwise stated.

Limitations

The approach chosen in this paper has several limitations. Firstly, the analysis only contains players drafted between 1990 and 2010 (inclusive), which may limit generalization to more recent years. Second, and probably the largest limitation, is that the majority of predictors consist of average career stats, which aren't available during a draft. This limits the analysis to the exploratory rather than the predictive realm. Another limitation of the analysis are factors that aren't captured by data. People that had injuries, greater positional value, less playing time, retired early, etc. are unaccounted for, since there's no feasible way to control for the slew of random kinks in the data. Hence, with only a little over 1000 observations, the analysis is limited in that regard. Finally, this paper only goes over one 80-20 data partition, of which there are many. Performance may vary across splits.

Results

This section contains the results of the previously described analysis performed on each of the two proposed models. It starts with a check of all necessary assumptions for inference upon a linear regression model. After the assumptions check, both models are analyzed and a prediction comparison is performed to address the central research question.

Assumptions

The assumptions for performing valid inference on a Normal error multiple regression model are as follows:

1. Linearity: The relationship between variables must reasonably linear. With multiple linear regression, this is harder to determine without domain knowledge, but it's not unreasonable to assume basketball metrics and career longevity could be linearly associated to some degree. The residuals plots for both models look fairly well distributed, but there's a visible diagonal line where the points won't cross 0. It's not enough to warrant a transformation, but would likely limit how useful a linear model can be. See the plots folder in the repository for exact plots.
2. Independence: Since the data consists of one basketball player per observation and each player is unique, the observations are assumed independent (see **Data** section).
3. Normality of errors: The residuals for both models should be approximately normally distributed. Note the QQ plots for both models appear to deviate at the ends, indicating some lack of normality. However, the usable region towards the center has the majority of points on the QQ line. Again, find diagnostic plots for both models in the plots folder of the repository.

4. Equal variances: The variance of residuals for both models should be relatively constant. Note there does appear to be some behavior indicative of unequal variances in the residual plots for both models. The variance starts small and grows to a stable amount, so the small variance at the beginning of the plot could possibly be attributed to lack of observations. See the residuals plots for both models in the plots folder.

5. Multicollinearity: Multiple highly correlated predictors can ruin MLR models. However, since the purpose of the following analysis is to determine how a model with almost certain multicollinearity (full model) performs against a model with much less (LASSO model), this assumption can be disregarded. Although clever variable selection with good domain knowledge can mitigate multicollinearity, there are many aspects of basketball that are correlated in general (see **Data** for more information).

Overall, the models would not fulfill assumptions for this dataset to produce reliable predictions. However, since this paper aims to compare the models with each other, the purpose of creating such models can still be fulfilled despite the failure of assumptions.

Model 1: Full model

Fitting an MLR model with all discussed predictors yields the following summary, created using Stargazer (Hlavac 2022):

The hypotheses for the test on each slope coefficient β_i were:

- $H_0 : \beta_i = 0$ (There is no linear relationship between predictor i and years played)
- $H_a : \beta_1 \neq 0$ (There is a linear relationship between predictor i and years played)

As can be seen from the table, the regression revealed multiple highly significant predictors within the full model: `mins_per_game`, `rebounds`_per_game,assists_per_game, `fg_percent`, `draft_cat` (the Late-second category) and `three_pt_percent`. This p-value is for the t-test of the slope coefficient and is less than the common alpha level of 0.05, leading us to reject H_0. Find a discussion of the pertinent assumptions for such a test in the **Assumptions** section. The fitted regression equation (without any variable selection) is

Table 1: Regression of all predictors on years played

	Dependent variable:
	years_played
mins_per_game	0.338^{***}
	(0.035)
pts_per_game	-0.085^{*}
	(0.051)
rebounds_per_game	0.535^{***}
	(0.086)
assists_per_game	0.219^{**}
	(0.102)
fg_percent	3.644^{***}
	(1.257)
three_pt_percent	2.369^{***}
	(0.783)
ft_percent	0.510
	(0.709)
draft_catMid-first	0.073
	(0.256)
draft_catEarly-second	-0.512^{*}
	(0.287)
draft_catLate-second	-0.750^{**}
	(0.334)
Constant	-2.329^{***}
	(0.705)
Observations	1,026
R^2	0.667
Adjusted R^2	0.663
Residual Std. Error	2.912 (df = 1015)
F Statistic	202.958^{***} (df = 10; 1015)
Note:	*p<0.1; **p<0.05; ***p<0.01

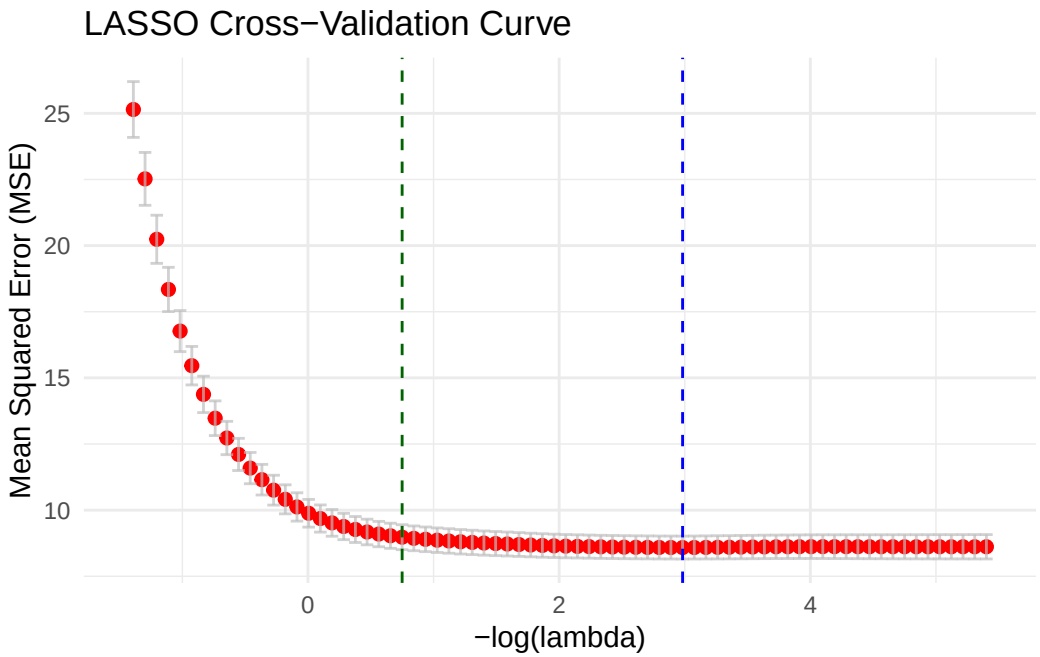
$$
\begin{aligned}
\text{years\_played} = &-2.329 \\
&+ 0.338(\text{mins\_per\_game}) \\
&- 0.085(\text{pts\_per\_game}) \\
&+ 0.535(\text{rebounds\_per\_game}) \\
&+ 0.219(\text{assists\_per\_game}) \\
&+ 3.644(\text{fg\_percent}) \\
&+ 2.369(\text{three\_pt\_percent}) \\
&+ 0.51(\text{ft\_percent}) \\
&+ 0.073(\text{draft\_cat}_{\text{Mid-first}}) \\
&- 0.512(\text{draft\_cat}_{\text{Early-second}}) \\
&- 0.75(\text{draft\_cat}_{\text{Late-second}})
\end{aligned}
$$

Beyond statistical significance, coefficient magnitudes provide insights to which predictors are more influential. The positive coefficients for `mins_per_game`(0.336), `rebounds_per_game` (0.526), and `assists_per_game` (0.242) suggest that players who have more playing time and contribute across such categories tend to have longer careers. Shooting efficiency especially appears to be more consequential than scoring volume. `fg_percent` (4.078) and `three_pt_percent` (2.855) have large positive coefficients, while `pts_per_game` has a non-significant negative coefficient, likely indicating some multicollinearity.

The draft position coefficients reveal a graded relationship: compared to lottery picks (the reference category), mid-first round picks show minimal difference (0.105 years), early-second round picks average 0.454 fewer years, and late-second round picks average 0.745 fewer years - with only the latter difference reaching statistical significance. This pattern suggests that the draft position advantage diminishes after the first round.

Model 2: LASSO selected model

To address multicollinearity and identify the most important predictors, LASSO regression was employed. For more information on how LASSO selects predictors, see the **Methods** section. Through 10-fold cross-validation on the training data, the optimal regularization parameter was determined.



The above figure displays cross-validation results for LASSO regression. The red dots represent the mean cross-validated error (MSE) for each λ value. The vertical bars (confidence brackets) show 1 standard error of the MSE estimates. The green dashed vertical line indicates λ_{min} (minimal MSE), and the blue dashed line indicates λ_{1se} (largest within 1 standard error of the minimum, selected for the model).

The cross-validation yielded a minimum mean squared error at $\lambda_{min} = 0.0325$. However, λ_{min} still kept all predictors, so the more parsimonious one-standard-error $\lambda_{1se} = 0.5303$ was utilized to build the LASSO model. The parsimonious LASSO model summary is as follows:

Table 2: Regression of LASSO selected predictors on years played

	Dependent variable:
	formula_str
mins_per_game	0.399^{***}
	(0.016)
rebounds_per_game	0.379^{***}
	(0.069)
Constant	-0.875^{***}
	(0.230)
Observations	829
R^2	0.671
Adjusted R^2	0.671
Residual Std. Error	2.905 (df = 826)
F Statistic	843.577^{***} (df = 2; 826)
Note:	*p<0.1; **p<0.05; ***p<0.01

The LASSO regression with λ_{1se} kept 2 of the 8 original predictors. It only kept `mins_per_game` (0.399) and `rebounds_per_game` (0.379), both with high significance levels. As can be seen, the adjusted R^2 is only 0.09 below that of the full model. The LASSO fitted model is simply denoted as

$$\text{years\_played} = -0.875 + 0.399(\text{mins\_per\_game}) + 0.379(\text{rebounds\_per\_game})$$

Prediction and Model Comparison

To evaluate the performance of the full model against the LASSO model, holdout validation was used. Both models were trained on a random 80/20 split of the overall NBA dataset, with 829 observations in the training set and 197 observations in the test set for the specified seed. Find a table of model performance metrics below. Lower RMSE, MAE, AIC, BIC, and Bias is better while higher R^2 and adj. R^2 is better.

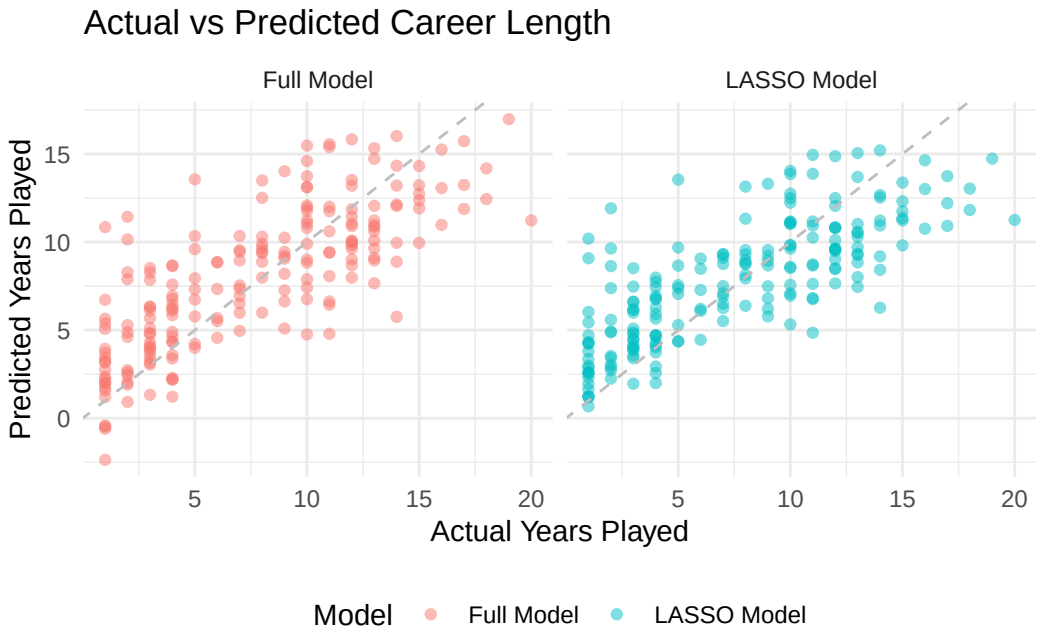
Table 3: Model Performance Metrics

	RMSE	MAE	R^2	Adj. R^2	AIC	BIC	Bias
Full Model	3.121	2.467	0.584	0.562	4109.7	4166.3	-0.394
LASSO Model	3.140	2.472	0.579	0.575	4125.5	4144.3	-0.211

Both models appear to have performed similarly. As expected, the full model did better than the LASSO model on many different metrics: Root Mean Squared Error, Mean Absolute Error,

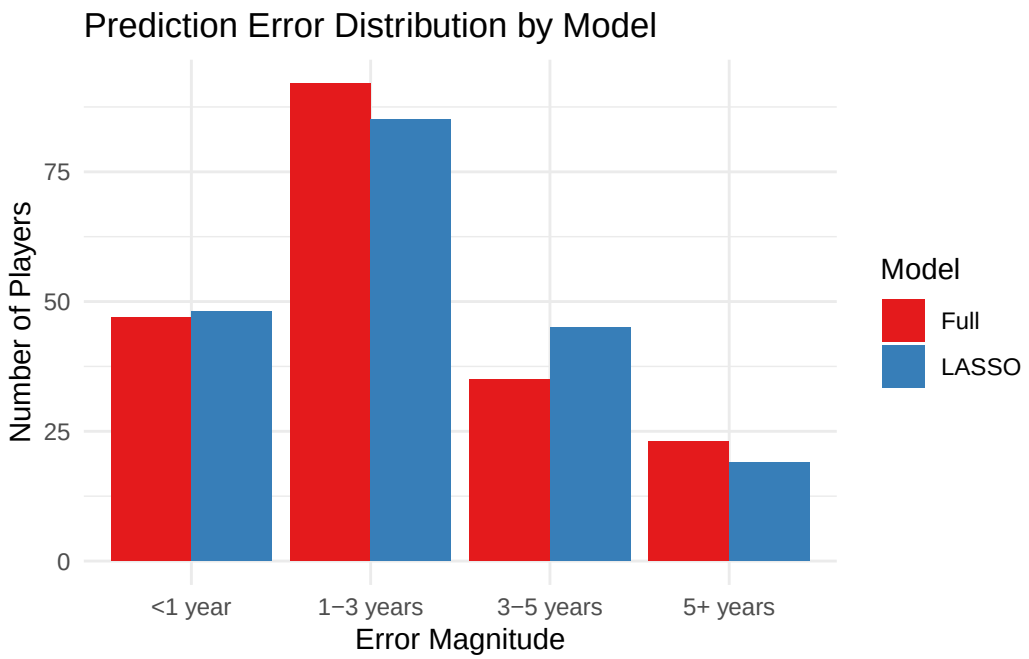
R^2 and AIC. However, the adj. R^2, BIC, and Bias are both poor, meaning the marginal gains it has over the LASSO model in other categories are likely due to the sheer number of extra predictors.

The following plot details the actual vs. predicted career length for both models, with the full model in red, LASSO model in blue, and the dashed lines being the prediction lines:



As can be seen, both models have reasonable predictions, as good as can be expected from a linear model. However, the tail ends (denoting extreme cases of low / high years played) are definitely weak points for both models.

Taking a look at the following plot, with the full model in red, the LASSO model in blue, and the error magnitude categories on the x-axis displays more interesting diagnostics.



The models have the majority of their predictions within 3 years of the actual values. However, the performance varies, with the LASSO model having a few more predictions within a year than the full model, but the full model having a lot more predictions within the 1-3 year error range than the LASSO model. It's difficult to tell from plot which model is more inaccurate since they're relatively even.

Conclusions

The analysis in this paper compared two regression models for predicting NBA career longevity. The full model used all eight proposed predictors, while the LASSO model selected only two using the tuning parameter that maximized parsimony: `mins_per_game` and `rebounds_per_game`. Both models performed similarly on test data, with the full model only having slightly better RMSE and R^2 and the LASSO model having better BIC and marginally more predictions than the other model within one year of the actual value.

The key finding from the analysis is that for this data, a simple two-predictor model explains nearly as much variance as the full model, implying that playing time and rebounding are the most important factors for career length (that are available in this data). Shooting efficiency was significant, but high scoring volume was not.

The study has glaring limitations, of course. Career stats that aren't known at draft time can't be used to create a predictive model. Lacking data on external factors also left the analysis blind to things that would have definitely impacted career longevity (injuries, position, etc.).

For future extensions, rookie stats would be a great place to start predictive modeling. Adding positional data and players' physical stats would also improve models. Despite the limitations of this project, the analysis shows that simpler models can be almost as effective as complex ones for explaining career longevity.

References

Hlavac, Marek. 2022. *Stargazer: Well-Formatted Regression and Summary Statistics Tables*. Bratislava, Slovakia: Social Policy Institute. https://CRAN.R-project.org/package=stargazer.

J, Friedman, Hastie T, and Tibshirani R. 2010. *Regularization Paths for Generalized Linear Models via Coordinate Descent*. https://doi.org/10.18637/jss.v033.i01.

Müller, Kirill. 2025. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Vivian Johnson, NCAA, and Ben Wieland. 2024. "NBA Drafts Between the Years 1990-2021." https://data.scorenetwork.org/basketball/nba_draft_1990-2021.html; SCORE Sports Data Repository.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.