NBA ROOKIE PERFORMANCE AND CAREER LONGEVITY: AN INFERENTIAL STUDY

Alex Baker, Kayla Manning, Matt Sheridan

Abstract

As players of one of the more fast-paced and injury-prone sports, the careers of NBA players can take unpredictable turns. However, success does not happen by chance; players cultivate fundamental skills over the course of their careers. Can statistics from a player's first season tell us anything about their career longevity? This analysis will assess that question, quantifying first-year performance and career longevity with rookie season statistics and whether the player lasts at least five years in the league. Our inferential study seeks to provide a better understanding of how rookie performance relates to career longevity in the NBA, and our findings could help coaches and fans recognize valuable player metrics early on in players' careers.

1 DATA

The provided dataset includes 20 columns total: a column with player names, a column that indicates whether or not a player lasted five years in the league, a column with games played in their rookie season, and 17 columns with per-game statistics. The `name` column will not provide any substance to our analysis aside from player identification, and the `surv_5yrs` column will serve as our outcome variable. This leaves us with 18 predictors from which to build our models. While we could build additional predictors by multiplying our per-game metrics by `games_played`, our model selection process will consider interaction terms that will serve essentially the same purpose.

1.1 Missing data

Missing data does not pose a major problem with this analysis. Only ten observations out of 1309 have any missing data at all, and this missingness only occurs in the three-point percentage variable. Observations with missing three-point percentage data all have zero attempted three-point shots per game, which leads us to believe these players never attempted a three-point shot during their rookie season.[1] To preserve these observations, we imputed these missing values with zero. While this may marginally reduce the effect size, making this change for only ten observations will have little, if any, impact on the results of our analysis.

[1] Upon further inspection, we recognized that many more players than ten players have zero three-point attempts per game. Because this metric is a per-game average, it is likely that the players with zero `t3p_attempts_per_game` but non-missing three-point percentage did attempt very few three-point shots. Most of these players with zero attempts per game have three-point percentages of 100, 50, 33.33, 25, and 0, which lead us to believe that these players with zero attempts per game but non-missing percentages took four or fewer shots in their first year. As a result of taking so few shots, the `t3p_attempts_per_game` rounds to zero.

1.2 Duplicated data

There are 22 duplicated names in the data. However, we found it hard to believe that there were truly six unique Charles Smiths, four unique Dee Browns, and two unique Tim Hardaways in the league, for example. Upon further investigation, some players – like the two Bobby Jones observations – had the same value for every predictor except for whether they lasted 5 years in the league. On the other hand, other duplicate names – such as the two Bob Martins – had entirely different first-year statistics and their fairly common names could reasonably belong to unique individuals. Because there were only 19 observations with duplicate names and statistics, we elected to research the individual names and only keep the correct observations. Compared to uniformly dropping duplicates or attempting to adjust modeling techniques, this research-based approach will provide more accurate insights into player longevity.

1.3 Train-test split

Because we aim to assess how rookie year performance relates to career longevity, we want our data to generalize across the NBA rather than the specific players in our dataset. This generalizability will allow for more accurate inferences. To assess model performance through this lens, we split our data into a train set for model building and a test set for model evaluation, using a 75% and 25% train-test split.

2 MODELS CONSIDERED

Our models will use `surv_5years`, a binary variable for whether a player lasts five years in the league, as the outcome variable. Originally coded as "Yes" and "No", we recoded the column as 1 and 0 to allow for logistic regression. We fit all of the below models as logistic regressions with the plan of updating models if we found evidence of poor fit.[2]

We began our model-building process with two baseline models: one that included independent main effects for all predictors, and another that included all predictors and their interactions. Then, we performed stepwise selection on the full interaction model, considering both AIC and BIC criteria.[3] Because the greedy algorithm behind stepwise selection does not always find the globally optimal model, we also adopted an all-subsets-considered approach using `glmulti` with the BIC criterion.[4] Finally, we used the predictors resulting from BIC stepwise selection to create a generalized additive model (GAM) to consider the possibility of nonlinearities.[5]

[2]If we observed overdispersion, we could have adapted our models to adjust for this with a beta-binomial fit. However, this was not necessary since the ratios of our deviances to our degrees of freedom were all approximately 1.

[3]Because the interaction model also includes main effects, we did not perform stepwise selection on the full main-effects-only model since the interaction model would retain any important main effects. We built the main-effects-only initial model for the all-candidates evaluation since the full interactions model was too large for glmulti to fit within a reasonable amount of time. Note that our stepwise selection used forward and backward selection combined.

[4]Due to limited time and computational resources, our candidate model approach used our full main-effects model rather than our full interaction model as a starting point. We opted to use the BIC criterion because it generally produces fewer predictors and we did not want to overfit the data.

[5]We elected to fit our GAM on these predictors rather than a full model to prevent overfitting; we fit this GAM after recognizing that the stepwise BIC model had strong performance in cross-validation while having fewer terms than comparable models. Because GAMs are much more flexible than traditional GLMs, we were very mindful of overfitting in our approach.

This process resulted in seven different models. To evaluate model performance and generalizability, we computed deviance on the test set, accuracy on the test set, and deviance from 10-fold cross-validation. To assess the fit of our binomial models, we also calculated the ratio of deviance to degrees of freedom to assess model dispersion. The below table displays these diagnostic measures:[6]

Table 1. Comparing model performance

	test_dev	test_accs	crossval_dev	dev_ratio
GamModel	2597.3155	0.7125	-	-
Initial Model	2624.9465	0.7095	107.1338	1.0699
Interaction Model	2928.8878	0.6758	170.7745	1.075
Stepaic	2707.8396	0.7003	104.9154	1.0488
Stepbic	2624.9465	0.7095	106.4301	1.0946
Glmulti aic	2652.5776	0.7064	106.7471	1.085
Glmulti bic	2652.5776	0.7064	107.0816	1.085

All of the models had a ratio of deviance to degrees of freedom under 1.1, which indicates that our binomial models do not suffer from overdispersion and are appropriate for this analysis. We also produced plots of the model residuals versus fitted values and Cook's distances, all of which looked appropriate across all models. That is, the residuals displayed uniform scatter and all of the Cook's distances were below 1, indicating that the model does not have a lack of fit and there are no overly influential points.[7]

2.1 Selecting our "best" model

Of these models, we will perform our inference based on our stepwise BIC model with no interactions. We selected this model for two main reasons: the simpler model contains fewer terms, and the benefits of BIC as opposed to AIC for our inferential purposes.

First and foremost, our stepwise BIC model exhibited strong performance in cross validation. The stepwise BIC model had the second-best cross-validated deviance, following only the stepwise AIC model. Of these two models, the stepwise BIC model has far fewer terms, with only three predictors: games played, offensive rebounds per game, and free throw percentage. This stands in contrast to the stepwise AIC model with 11 main effects terms and 5 interaction terms. A variance inflation test of the predictors in stepwise AIC model reveals that several of the terms are highly collinear, which does not pose an issue for the stepwise BIC model, as demonstrated in Tables 2 and 3.

By convention, a variance inflation factor exceeding 10 provides evidence of multicollinearity. Given that over half of the predictors from our stepwise AIC model meet this criterion, any inferences drawn from this model would be very imprecise due to the inflated coefficients and standard errors resulting from collinearity. Following the logical principle of Occam's razor – which posits that simpler models

[6]Note that although the table displays the test deviance and accuracy, we performed model selection based on the cross-validated deviance. We only used the test set to evaluate the final performance of the model and to re-fit the model for our inferences.

[7]See the Appendix for the diagnostic plots for each of the fitted models.

Table 2. Variance inflation factors for stepwise AIC model

variable	.
games_played	5.679806
min_per_game	12.256287
fg_made_per_game	29.454207
t3p_made_per_game	33.667017
t3p_attempts_per_game	35.531707
t3p_percent	1.854233
ft_percent	3.085923
oreb_per_game	3.627041
assists_per_game	3.407794
steals_per_game	65.435276
blocked_shots_per_game	27.200585
games_played:fg_made_per_game	33.584318
ft_percent:steals_per_game	68.611790
min_per_game:steals_per_game	19.374228
fg_made_per_game:blocked_shots_per_game	6.341112
games_played:blocked_shots_per_game	28.794789

Table 3. Variance inflation factors for stepwise BIC model

variable	.
games_played	1.105665
oreb_per_game	1.146073
ft_percent	1.105089

are often better – we selected the more parsimonious stepwise BIC model.

While the absence of collinearity already makes our stepwise BIC model more attractive than our stepwise AIC model, the inferential purposes of this analysis make BIC stepwise selection more attractive than AIC model selection, not to mention that BIC is more conservative in its relationship estimations. BIC is consistent and asymptotically efficient for parametric models, making it appropriate when inference is the main goal. On the other hand, AIC is not asymptotically efficient for parametric models and can converge to the wrong model with $n \to \infty$.[8]

[8] AIC is asymptotically efficient for a non-parametric framework and is more appropriate when prediction is the main goal.

3 CONCLUSIONS FROM THE CHOSEN MODEL

While not perfect, our chosen model correctly classified 70.95% of test observations, indicating that it provides some valuable information about how rookie-year metrics relate to career longevity.[9] Our stepwise selection process with the BIC criterion kept three final predictors in the model – games played, offensive rebounds per game, and free throw percentage – all of which are significant at the $\alpha = 0.05$ significance level. The below output displays our selected model, re-fit to the test data:[10]

Table 4. Stepwise BIC model output

	Dependent variable:
	surv_5yrs
games_played	0.038^{***}
	(0.009)
oreb_per_game	0.674^{***}
	(0.232)
ft_percent	0.015
	(0.013)
Constant	-3.424^{***}
	(0.955)
Observations	327
Log Likelihood	-186.694
Akaike Inf. Crit.	381.389
Note:	*p<0.1; **p<0.05; ***p<0.01

Notably, the number of games played in a player's rookie season serves as a significant predictor in nearly all of our models, including our stepwise BIC model. Holding offensive rebounds per game and free throw percentage constant, each single unit increase in games played in the rookie season is associated with a multiplicative increase of 1.039 in the expected odds of lasting at least five years in the NBA. Intuitively, coaches tend to play players with the potential to perform well on the court, and strong performance should reasonably translate into a strong career.

In addition to games played, our stepwise-selected BIC model deemed offensive rebounds per game and free-throw percentage as predictors worth keeping in our final model. Substantively, the inclusion

[9] Our GAM model had the best performance on the test set, but it also lacks interpretability in the inferential context of this analysis. While it might perform best in predictive settings, we will focus on the stepwise BIC model for most of our discussions surrounding rookie performance and career longevity.

[10] If we use the same data to select a model and then to make inference from it, the model will produce overly optimistic predictions. Because we conducted our stepwise selection on the train set, we refit the data on the test set to perform inference.

of these variables makes sense since they reflect both soft and hard skills of players.

For each additional offensive rebound per game, the expected odds of a rookie lasting at least five years in the league is multiplied by 1.962 when holding the other variables constant. Because defense typically stands between offensive players and the rim, players of all positions must work to get offensive rebounds. As a result, a player who successfully gets offensive rebounds likely possesses a great deal of grit, tenacity, court vision, and confidence. Tall centers must work their way around their tall opponents under the basket. On the other hand, smaller guards and forwards must have the speed to maneuver around their opponents. Tall players have slightly less of an advantage in offensive rebounds compared to defensive rebounds since the defenders focus less on boxing out smaller offenders. From a logistical standpoint, offensive rebounds provide a team with extra chances to score points in the same play, can translate into potential free throw opportunities, and have the power to frustrate the defense. With all of this in mind, it is no surprise that a player with the personal and physical qualities to get offensive rebounds would be a valuable mainstay on an NBA roster, as determined by our model.

A single percentage point increase in rookie-season free throw percentage is associated with a multiplicative increase of 1.016 in the expected odds of lasting at least five years in the league, holding other variables constant. Similar to offensive rebounds, a strong free throw percentage also reflects elements of character and physicality that reasonably translate to career longevity. Free throws are just that: free, uncontested, single-point shots granted after shooting fouls or in the bonus period. A strong free throw percentage reflects disciplined practice[11] and the ability to perform under pressure. In addition to reflecting desirable elements of a player's character, free throws score points. A reliable free throw shooter is often a reliable scorer, and reliable scorers are the type of players who should last at least five years in the league.

To account for nonlinearities in the predictors selected by our stepwise BIC model, we fitted a generalized additive model to supplement our generalized linear model. All three smoothing terms were significant at the $\alpha = 0.05$ level, and we see some nonlinearities in all of the below smooth plots.[12] The effect of offensive rebounds per game on career longevity increases at higher levels of offensive rebounds, as does the effect of games played. We see that free-throw percentage has a decreasing effect on the log odds of lasting five years for very low values, but the effect of free throws increases from about 60% and onward.[13] Across all variables, there is much greater variability at the extreme ends of the spectrum, with far fewer low observations for games played and free throw percentage and far fewer high observations for offensive rebounds per game:

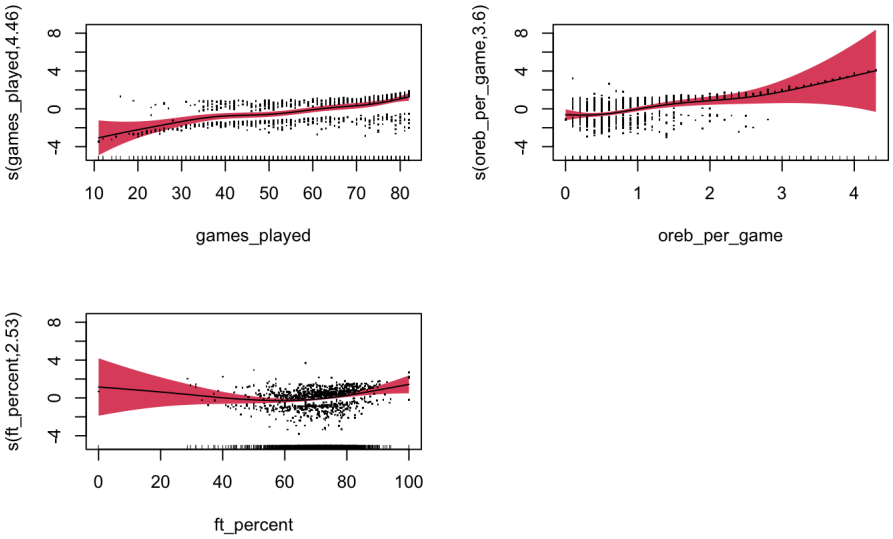
[11] Interestingly, NBA teams used to have dedicated free throw coaches. However, teams have gotten rid of these coaches over the past five seasons, which has resulted in a decline in overall free throw percentages. `https://www.sportico.com/leagues/basketball/2021/nba-free-throw-rates-1234647427/`

[12] In addition, a likelihood ratio test between the GAM model and the GLM model with the identical predictors was significant with a p-value of approximately 0.01. This further indicates that there are non-linearities in the data.

[13] While a high free throw percentage could result from only taking a few shots, strong free throw percentages – typically in the high 80's or above – are a valuable asset and serve as a significant predictor of future success in our model. This could be the source of some discrepancy in the GLM's predictions, as the model has no way of discriminating between players who have few shots taken.

Table 5. GAM model output

	edf	Ref.df	Chi.sq	p-value
s(games_played)	4.463232	5.479330	86.44533	0.0000000
s(oreb_per_game)	3.597132	4.461449	43.72568	0.0000000
s(ft_percent)	2.526495	3.296088	15.88747	0.0017932



3.1 Summary of approach

Prior to conducting any analysis, we split our data into a train and test set. We sought to perform inference rather than prediction, but we adopted a train and test split to ensure that our inferences generalize across the NBA. Using our training set, we fit seven models before selecting our "best" model. We evaluated the performance of each of these models using the cross-validated deviance. With the model performance, the number of predictors, and properties of the BIC in mind, we selected our stepwise BIC model as the model of focus for our inferences. This model indicates that games played, free throw percentage, and offensive rebounds per game are important predictors of whether a player lasts at least five years in the NBA.

Very few players are truly well-rounded, and even household names like Michael Jordan, Lebron James, and Kevin Durant underperform occasionally. Our final stepwise BIC model considers that players specialize and emphasizes fundamentals: games played, free throw percentage, and offensive rebounds. By not including many of the other position-specific, less fundamental variables – such as blocks or three-point percentages – our model tolerates player-specific weaknesses. However, our model also demonstrates tolerance of weaknesses in the included variables.[14]

[14]See the Appendix for a demonstration of model tolerance for low free throw percentage, low offensive rebounds, and incorrectly coded outcomes.

3.2 Limitations

3.2.1 Model limitations

Non-linearities in the selected predictors pose a key limitation in our model. The generalized additive model with the predictors in our stepwise BIC model had the highest test accuracy, indicating that the model performs well in a predictive setting and should in theory be generalizable. However, we sought to *understand* the relationship between rookie season statistics and career longevity. It is much more challenging to conceptualize the relationship between games played and career longevity as somewhere between quartic and quintic, and the same idea applies to the offensive rebounds per game and its cubic-quartic hybrid and the free throw percentage and its quadratic-cubic hybrid. We did not see the 0.3% increase in test accuracy as worth the loss of interpretability by focusing solely on the GAM. Rather, we focused on the predictors that yielded two of the top-performing models and interpreted the substantive meaning behind the models.

3.2.2 Data limitations

Obviously, rookie season metrics provide an incomplete look at factors explaining career longevity. Our model would benefit dramatically from the inclusion of more variables. Predictors such as height, age, draft rank, and college recruitment stats would shed more light on predictors leading to lasting NBA careers. For example, people above seven feet tall are approximately 10% more likely to play in the NBA than those under seven feet.[15] The inclusion of more player characteristics would preserve the focus on the rookie season while adding more nuance to the analysis.

3.2.2.1 No time-specific variables In addition to basic player recruitment information, a time variable would provide immense value to our analysis. From a historical standpoint, the number of games played in a season has evolved with time, with the standard 82-game season beginning in the 1967-1968 season.[16] In addition to changes in the number of games, new rules – such as the addition of the three-point line in 1979 – have changed the style of play.[17] The provided data does not provide the year for each observation, so we cannot control for these time-related changes in rules or league structure.

On a more narrow timescale, many of the players are currently in the league and had their rookie season fewer than five years from data collection. As a result, players who will last five years in the league are coded as not lasting five years. This will decrease the effect size of important variables, which will therefore decrease the power of our analysis.[18] Unfortunately, the time constraints of this project limited our ability to conduct extensive searches of each of the 1309 players' draft years and contracts, and the lack of a year variable means that we cannot easily control for players who had their rookie season less than five years prior to data collection.

[15] https://www.bostonglobe.com/metro/regionals/west/2014/03/09/footers-percent-chance-playing-nba/fNnbP8zybYfXZtsw0eYPDK/story.html#:~:text=From%20this%2C%20he%20further%20deduced,someone%207%20feet%20or%20taller.%E2%80%9D

[16] https://www.foxbusiness.com/sports/nba-82-game-regular-season-schedule

[17] https://www.usab.com/youth/news/2011/06/the-history-of-the-3-pointer.aspx#:~:text=The%20league%20didn't%20adopt,following%20suit%20a%20year%20later.

[18] Fortunately, our model actually demonstrates accurate predictions in the face of inaccuracies. See the Appendix for an illustrative example of this with Karl Anthony Towns, who is incorrectly coded in the data as not lasting five years in the league.

3.2.2.2 Rookie-season injuries In addition to time-related variables, the lack of injury-related variables limits our ability to gauge career longevity with rookie year data. If we sought to maintain the rookie-season focus of our analysis, a binary variable indicating whether a player suffered an injury in their rookie season would provide great value. For example, Markelle Fultz battled a rookie-season injury for three years, which limited his ability to perform well in season-long metrics and per-game averages.[19] He is now a star point guard and consistent starting player with an improving contract. These "virtual guarantee" injuries extend the time a player will be in the league simply because it takes time to recover. However, we have no way of measuring this with the provided data.

3.2.2.3 Post-rookie data A model of career longevity would benefit even further with an expansion of focus beyond the rookie season. First and foremost, statistics over a longer timeline would provide more information about a player's performance beyond the first year. Additionally, injuries can occur at any time in a player's career, so a measurement of injuries and their timing between years one and five would provide much better context for player statistics and outcomes. Also, a measurement for trades and team chemistry would provide more context about a player's team during their first five years. Finally, a variable that measures whether players continue their careers overseas would change the scope of the study question but would capture a more thorough timeline of players' careers.

3.3 Final comments

Many factors could determine whether a player lasts five years in the league, and rookie performance only plays a narrow role in a player's overall career. While far from perfect, our chosen model correctly classified nearly 71% of observations in the test set, which indicates that it provides generalizable information about how rookie performance relates to career longevity in the NBA. From this analysis, we see that games played, free throw percentage, and offensive rebounds per game have a strong relationship with a player's potential to play at least five years in the league. While the addition of new variables might have strengthened our analysis and made our predictions more position-specific, these findings indicate that players should not lose sight of fundamentals.

[19]Markelle Fultz is not included in our dataset, but we cited him here for illustrative purposes since he is a rather well-known player plagued by this exact injury delay.

4 APPENDIX

4.1 Tolerance of weaknesses

4.1.1 Player weaknesses

By highlighting player fundamentals rather than focusing on position-specific variables, our model tolerates player weaknesses. However, our model demonstrates strong performance even when great players have lower-than-expected values for one of the included variables. For example, Shaquille O'Neal had a rookie free throw percentage of only 59.2%, which places him in the bottom 14% of observations by that metric. However, he ultimately played 19 seasons in the league.[20] Despite his very low free throw percentage, the model correctly predicts that he lasted five years in the league. Similarly but to a lesser extent, Steph Curry averaged 0.6 offensive rebounds per game during his rookie season, which places him in the bottom 40% of players. However, the model still correctly classifies him as lasting at least five seasons in the league.

Table 6. Selected player predictions

name	games_played	ft_percent	oreb_per_game	surv_5yrs	predicted_prob
Stephen Curry	80	88.5	0.6	1	0.8288555
Karl-Anthony Towns	82	81.1	2.8	0	0.9689870
Shaquille O'Neal*	81	59.2	4.2	1	0.9846279

4.1.2 Data weaknesses

Finally, the model demonstrates tolerance for observations that may be incorrectly coded in the data. As mentioned in the main analysis, some players had their rookie season fewer than five years prior to data collection. In these cases, players did not have the chance to play for five years and are therefore coded as not surviving five years. Karl-Anthony Towns, for example, was drafted in 2015 and has played in the league since,[21] but the dataset lists him as not lasting five years. Despite being fit on the incorrectly coded observation, our model correctly predicts that Karl-Anthony Towns does last at least five years in the league, as displayed in Table 6. Overall, the model tends to produce correct predictions at a greater rate for players who played at least five years in the league. Considering that a subset of those coded as not lasting five years will ultimately do so, some of the predictions counted as incorrect may actually be correct.

Table 7. Predictive accuracy by outcomes across entire dataset

surv_5yrs	accuracy
0	0.5501022
1	0.8268293

[20] https://www.statmuse.com/nba/player/shaquille-o'neal-2780#:~:text=Center%20Shaquille%20O'Neal%20played,Hall%20of%20Fame%20in%202016.

[21] https://www.basketball-reference.com/draft/NBA_2015.html

4.2 Diagnostic plots

The below plots visualize the residuals versus fitted values and Cook's distances for each of our models. As described in the main text, none of the plots raise any concerns regarding our models. The residuals are uniformly scattered and all of the Cook's distances are less than 1. With that, we do not have evidence of poor fit in any of our models.