Predicting Retired Player Jersey Numbers: Statistical Data Analysis Application

Eduardo Martinez

May 7, 2021

Introduction

Statistical methods are commonly used to forecast and predict future outcomes by training regression models using data (Hastie, Tibshirani, Friedman, 2001). For example, logistic regression or classification methods train models to learn and assign labels or classes to points in a data set. A model is trained effectively if the data used is larger in terms of number of rows; however, data sets that contain a large amount of hyperparameters weaken a model (Shar, 2017). This phenomenon is referred to as the curse of dimensionality (Hastie, Tibshirani, Friedman, 2001). If a data has many hyperpameters, subset selection methods should be applied to find the most crucial predictors.

Information criteria are used to find the best subset for Generalized Linear Model (GLM). McLeod & Xu (2010) and Zhang, (2016) describe several key goodness of fit criteria, including Alkaline Information Criterion (AIC) and Bayesian Information Criterion (BIC). All of these criteria seek to maximize the log-likelihood of a model while applying constraints when models have excessive predictors. Letting D denote the penalized form the of deviance, here are their formulas:

- $AIC = D + 2k$ where k is the number of predictors out of p predictors in the full model.
- $BIC = D + k * log(n)$ where n is the size of the data.

The model that minimizes AIC and BIC is the optimal subset. The value of AIC describes the estimated entropy of a model. Minimizing BIC leads to maximum posterior probability (McLeod & Xu,2010). Compared to the AIC, the BIC tends to favor smaller models. In fact, when $n > 7$, the BIC penalty is larger than the AIC penalty.

Miller (1984) provided a detailed descriptions of the theory regarding various methods. Also, modern explanations of best subset selection criterion can be found in "The Elements of Statistical Learning" (Hastie, Tibshirani, Friedman, 2001).

Background & Methods

The primary objective in this project was to predict whether an NBA players jersey number would be retired or not. To make such prediction, logistic regression analysis was applied to portions of multiple data sets (Basketball Reference, 2019; Goldstein, 2018; Land of Basketball, 2021; NBA Hoops Online; Weak Side Awareness, 2011)

This study was completed entirely using R with RMarkdown in RStudio. The main package used for the regression methods used in this study was "bestglm', whch required the 'leaps' package; see the accompanying citation for the package documentation (Mcleod, Xu, & Lai, 2020). McLeod & Xu (2010) and Zhang (2016) completed various example experiments using R, which were frequently utilized and used in this study.

The Data

The data obtained for this study tracked NBA statistics of players from 1982 to 2017 with each year representing an entire regular season (e.g, 1998 denotes the 1997-1998 NBA regular season).

Under the assumption that NBA player reach their prime (best performing years) around the middle of their career, players that retired before 1985 or began their career after 2014 would represent misleading data because their prime years would be excluded from the data. For example, Bob Lanier

played in the NBA for 14 seasons before retiring in 1984. His jersey number was retired by two teams. He averaged 20.1 points per game and 10.1 total rebounds per game during his career. However, this dataset only includes his last 3 seasons, for which he averaged only 13 PPG and 5.6 TRPG. Including Bob Lanier in the dataset would be misleading. To control for these instances, players that less than 4 years or less than 200 games were removed from the data set.

Additionally, players that were in a team for only one or two years usually do not make a meaningful impact worthy of getting their jersey number retired. The only NBA player that played with a specific team for two years or less and got his jersey number retired was Malik Sealy, whose number was retired primarily to honor him after he died in a car crash mid-season ("Retired Jersey Numbers". Since this is an extreme case, all players that played for a specific team for two years or less were removed from the data. When predicting whether a player's jersey should be retired or not, a players entire career should be considered, rather than each single seasons. Hence, the data was further condensed and split accordingly.

The final data utilized was separated into two data sets:

1. **Career Data**: Statistics and achievements (predictors) pertaining to the entire NBA career of 1248 players.
2. **Team Data**: If players in the Career Data played at least three seasons for multiple teams, their data was separated based on the statistics realized in each team. For instance, Kevin Garnett is listed twice in this data set because he played at least three seasons with the Minnesota Timberwolves and the Boston Celtics.

Both data sets consisted of the same players, although the Team Data contains 1798 rows of data compared to 1248 rows in Career Data.

The were 12 predictors considered:

1. Championships/Titles Won
2. All-Star Selections
3. Seasons played
4. Games Played (GP)
5. Games Started (GS)
6. Points Per Game (PPG)
7. Turnovers Per Game (TOV)
8. Player Efficiency Rating (PER)
9. Usage Percentage (USGp)
10. Win Shares (WS)
11. Box Plus/Minus (BPM)
12. Value Over Replacement (VORP)

Note, the Team Data did not include All-Star selections because a source distinguishing selection earned by a player with respect to each team could not be found.

For a detailed description of these variables, see the Glossary on Basketball Reference.

Is a player's jersey number retired?
The answer to this question was encoded using dummy variables: $0 = NO$ and $1 = YES$.

Train-Validation-Test Subsets

The data was split into three subsets: Training, Validation, and Testing. Shar, T. (2017) describes why it is useful to add a validation set instead of using only a train and test set.

The Test Set consisted of 276 out of the 1248 total players, representing 22.12% of Career Data, and 20.86 of Team Data. All players who played in 2017, which is the last year recorded in the data, were intentionally assigned to the Test Set. A players jersey number can only be retired after the player retires from the NBA. Hence, the true outcomes for most players in the Test Set is yet to be determined. Specifically, retired jersey number were obtained via data updated in 2019. Manu Ginobili, who retired from the NBA in 2018 and got his jersey number retired in 2019, is the last player that is correctly labeled in the data. Players that played in 2019 were separated into a set named "Unknown Test Set". The remaining players were assigned into a set named "Known Test Set." Since some outcomes were unknown, it was

essiential to add a validation set.

The remaining 1248 players were randomly split between the Train and Validated sets using a Train-Validation split ratio of 70/30. The Train and Validation set represented 54.49% and 23.4% of Career Data, respectively, and 56.56% and 22.58% of the Team Data, respectively.

Subset Models

There were 12 and 11 predictors considered in the Career and Team data, respectively. If every one of these predictors were used to predict whether a player's jersey number will be retired or not, the resulting model will be highly flexible. Consequently, it will result in overfitting the data (Miller, 1984). Such models tend to perform well on training data but poorly on testing data.

The accompanying R package utilized to find the best subset was 'bestglm' (McLeod, Xu, & Lai, 2020). A smaller model that does well on training and testing data is preferred. The best subset for each model was found using forward selection with a BIC. Note, AIC was also used, but some predictors in the best subset produced were not statistically significant *(p-value < 0.05)*. Therefore, BIC was the the selection method criteria used.

Initially, an *exhaustive* search was used, but since there were more than 10 predictors, an exhaustive searched took about a minute two run each time in R. Further, the best subsets found using an exhaustive search ended up being the same as step-wise forward search. Once this was realized, forward selection was used as it only took about 20-30 seconds to execute.

Subset Model - Career Train Set: Jersey_Retired = -6.179 + 0.295 * All-Star_Selections + 0.295 * WS

Subset Model - Team Train Set: Jersey_Retired = -9.072 + 0.006 * GP + 0.006 * WS

These models were used to predict the outcomes of the Validation Set. Then, the Train Set and Validation Set were combined into a new data set referred as the *Combined Train*

Set. The best subset was reevaluated and the resulting model was retrained via the same process but using the Combined Train Set.

Subset Model - Career Combined Train Set:

Jersey_Retired = -6.179 + 0.295 * All-Star_Selections + 0.295 * WS

Subset Model - Team Combined Train Set:

Jersey_Retired = -9.072 + 0.006 * All-Star_Selections + 0.006 * WS

For the Career Data, the updated model selected the same predictors as those in the training model, and their coefficients are nearly the same. In contrast, the updated model for Team Data selected the predictor number of seasons played (nSeasons) instead of GP, while the intercept and WS coefficient were marginally altered. At any rate, nSeasons and GP are co-linear as their correlation coefficient = 0.95.

These updated models were used to predict the Test Set (Known Outcomes) and Test Set (Unknown Outcomes). Although an accuracy score cannot be assigned to the Test Set (Unknown Outcomes), the model's predictions provide insight on which NBA players will get their jersey numbers retired in the future.

It is hypothesized that the Team Models will outperform the Career Models overall because jersey numbers are retired by a team and not the entire league.

Results

The confusion matrices for each model's prediction accuracy are expressed in table format throughout this section.

Train Set **(Table 1A & Table 1B)**

Both models did exceptionally well on the training data, achieving over 95% accuracy. The Team Model outperformed the Career Model by only 0.09% overall, but the Career Model did much better at predicting, which players got their jersey number. The Career Model correctly predicted 58.5% of the players that got their jersey number retired, while

the Team Model correctly predicted 40.7% correctly. Compared to there overall accuracy, the models did a poor job at predicting which players got their numbers retired.

Table 1A: Career Model		
Predictions	True Outcome	n
0	0	621
1	0	6
0	1	22
1	1	31

Train Set Accuracy = 95.88 %

Team Model		
Predictions	True Outcome	n
0	0	954
1	0	9
0	1	32
1	1	22

Train Set Accuracy = 95.97 %

both achieved over 90% accuracy. The Career Model's accuracy decreased by 2.04%, while the and Team Model's only decreased by 1.13%. When predicting which players had their number retired, the Career Model's accuracy dropped to 43.5%, which was not much greater than the 39.1% accuracy exhibited by the Team Model. For these outcomes, the Career Model's accuracy decreased by 15%, but the Team Model's accuracy decreased by only 1.05%. This suggests that the Career Model might have overfilled the training data.

Validation Set (Table 2A & Table 2B)

A models performance is evaluated on how well it predicts the validation set, since flexible models can always do well on training data. The models demonstrated they can also do well on the Validation Set as they

Table 2A: Career Model		
Predictions	True Outcome	n
0	0	264
1	0	5
0	1	13
1	1	10

Validation Set Accuracy = 93.84 %

Table 2B: Team Model		
Predictions	True Outcome	n
0	0	376
1	0	7
0	1	14
1	1	9

Validation Set Accuracy = 94.83 %

Combined Train Set (Table 3A & Table 3B)

The Combined Train Set produced a model that yielded accuracy scores over 95%. It is not alarming that their accuracy percentages dropped by less than 1% because the best subset found using the BIC method attempted to generalize the models so that they can perform consistently well on new data. In terms of each model's performance in predicting player's that got their jersey numbers retired, the Combined Career Model yielded 52.6% accuracy, and the Combined Team Model was 40.3% accurate. Among these set of players the Career Model's accuracy was 58.5%, 43.5%, and now 52.6%, performing inconsistently. By contrast, the Team Model percentages ranged between 39.1% to 40.7%, showing consistency albeit lower accuracy.

Table 3A: Career Model		
Predictions	True Outcome	n
0	0	885
1	0	11
0	1	36
1	1	40

Combined Train Set Accuracy = 95.16 %

Table 3B: Team Model		
Predictions	True Outcome	n
0	0	1334
1	0	12
0	1	46
1	1	31

Combined Train Set Accuracy = 95.92 %

Test Set (Known Outcomes) (Table 4A & Table 4B)

On the Test Set (Known Outcomes), the Team Model outperformed the Career Model by 1.81%. The Team Model's prediction actually increased marginally by 0.68% on the Test Set (Known Outcomes). Both models correctly predicted only 33% of the players

that got their jersey numbers retired. However, only 6 players had their numbers retired in this test set, which is not enough of a sample size to make meaningful conclusions.

Table 4A: Career Model		
Predictions	True Outcome	n
0	0	89
1	0	1
0	1	4
1	1	2

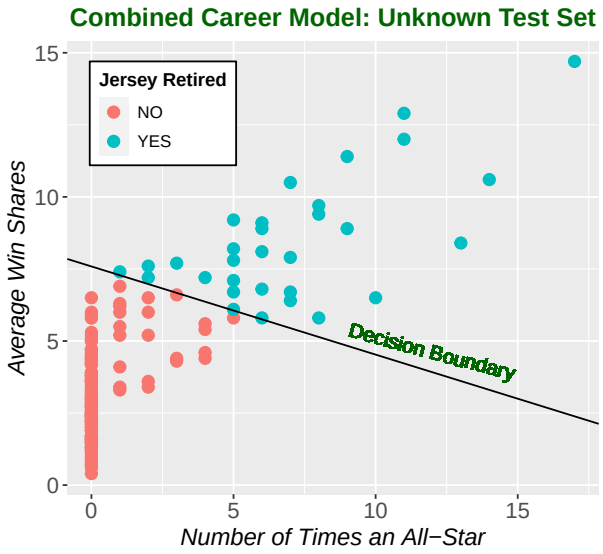
Known Test Set Accuracy = 94.79 %

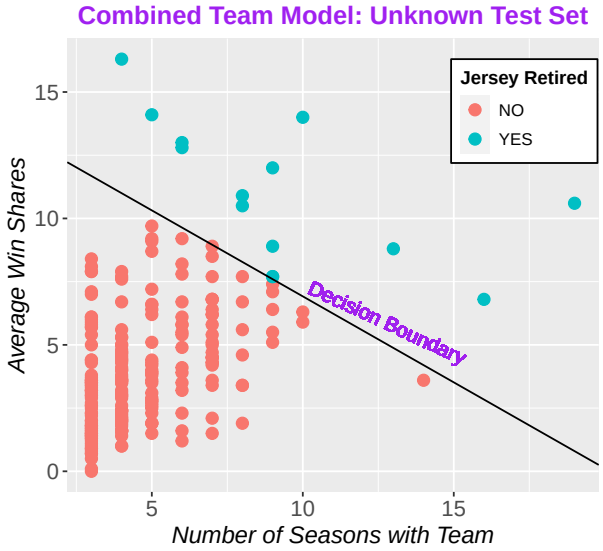
Table 4B: Team Model		
Predictions	True Outcome	n
0	0	140
1	0	1
0	1	4
1	1	2

Known Test Set Accuracy = 96.6 %

Test Set (Unknown Outcomes)

Since the best subset for both the Full Career Model and the Full Team Model only had two predicting variables, it is possible to visualize how these models classified values in the Unknown Test Set based on each player's entire career.





The slope of the decision boundary for the Career Model is relatively flat. This means that WS has a much greater affect on the predicted outcome than number of All-Star se-

lections. In fact, a player with only 1 All-Star Selection is predicted to have his jersey number retired. In comparison, the minimum WS among players predicted to get their number's retired is 6. According to the Decision Boundary, even 15 All-Star selections is not enough for a player to get their jersey number retired unless they have at least 3 WS.

In the Team Model, the slope suggests that both predictors have approximately equal influence on the outcome, but win shares has a slightly greater affect. The minimum number of seasons is 3, yet 4 seasons was enough to retire a players jersey number (top left point). Meanwhile, 7 win shares appears to be the minimum amount among the players that are predicted to get their jersey number's retired.

The two tables below are arranged in descending order by each player's (point on the graphs above) euclidean distance from the decision boundary of its linked model. In other words, the players listed at the top of the table are farthest away from the decision boundary, and the players listed last are closest to the decision boundary. This means that players listed last are least likely to get their jersey numbers retired compared to players listed at the top, although they are still predicted to get them retired.

Career Model: Unknown Test Set		
Player	All-Star	WS
LeBron James	17	14.7
Chris Paul	11	12.9
Kevin Durant	11	12.0
Dirk Nowitzki	14	10.6
James Harden	9	11.4
Stephen Curry	7	10.5
Dwyane Wade	13	8.4
Anthony Davis	8	9.7
Dwight Howard	8	9.4
Russell Westbrook	9	8.9
Damian Lillard	6	9.1
Blake Griffin	6	8.9
Kawhi Leonard	5	9.2
LaMarcus Aldridge	7	7.9

Remaining players not listed but included in the table above: Pau Gasol, Jimmy Butler, Carmelo Anthony, Kevin Love, Kyrie Irving, Tony Parker, Al Horford, Marc Gasol, Paul George, Paul Millsap, Vince Carter, Giannis Antetokounmpo, Rudy Gobert, Andre Drummond, DeAndre

Jordan, Kyle Lowry, Klay Thompson

Team Model: Unknown Test Set			
Player	Team	Seasons	WS
Dirk Nowitzki	DAL	19	10.6
LeBron James	CLE	10	14.0
LeBron James	MIA	4	16.3
Kevin Durant	OKC	9	12.0
Tony Parker	SAS	16	6.8
Dwyane Wade	MIA	13	8.8
James Harden	HOU	5	14.1
Chris Paul	LAC	6	13.0
Chris Paul	NOP	6	12.8
Dwight Howard	ORL	8	10.9
Stephen Curry	GSW	8	10.5
Russell Westbrook	OKC	9	8.9
LaMarcus Aldridge	POR	9	7.7
Marc Gasol	MEM	9	7.7

Notable Players Left Out

Contrary to the previous two tables, the players in following two tables are arranged in *ascending* order based on their euclidean distance. Players at the top were very close to meeting the minimum requirements set by the decision boundary. The list of players predicted to not get their jersey numbers retired is and 214 players for the Career and Team Data, respectively. Only the top 12 players and Giannis Antetokounmpo, who is ranked 52, are listed from the Team Data.

Giannis was added because he was the MVP in 2019 and 2020, but the last year tracked in the data was 2017. Interestingly, Giannis was predicted to get his number retired by the Career Model.

Career Model		
Player	All-Star	WS
Draymond Green	3	6.6
John Wall	5	5.8
Isaiah Thomas	2	6.5
DeMar DeRozan	4	5.6
Kemba Walker	4	5.4
Joakim Noah	2	6.0
DeMarcus Cousins	4	4.6
Luol Deng	2	5.2
Rajon Rondo	4	4.4
Bradley Beal	3	4.4
Derrick Rose	3	4.3
Khris Middleton	2	3.6
Victor Oladipo	2	3.4

Team Model			
Player	Team	Seasons	WS
Blake Griffin	LAC	7	8.9
DeAndre Jordan	LAC	9	7.4
Kawhi Leonard	SAS	6	9.2
Pau Gasol	LAL	7	8.5
Al Horford	ATL	9	7.1
Andre Iguodala	PHI	8	7.7
Udonis Haslem	MIA	14	3.6
Mike Conley	MEM	10	6.3
Anthony Davis	NOP	5	9.7
Luol Deng	CHI	10	5.9
Kyle Lowry	TOR	5	9.2
Joakim Noah	CHI	9	6.4
Giannis Antetokounmpo	MIL	4	6.7

Conclusion & Discussion

All models did were very accurate overall (over 90% accurate), especially at predicting when a player's jersey will not be retired. However, they did not predict when a player's jersey will be retired as well. A possible explanation for such poorer performances is that sometimes players get their jersey numbers retired for non-statistical reasons. For instance, some players like Malik Sealy and Reggie Lewis were honored after they passed away, unexpectedly; other players like Pete Maravich, who's number was retired by the New Orleans despite the fact he did not play for them, were honored for their impact on the community ("Retired Jersey Numbers".

It is difficult to say which model performed better between the Career Model and the Team Model. However, the main advantage of the Team Model was that its accuracy was very consistent on all subsets predicted and trained by. The Career Model did well in the training data, but its efficiency dropped when predicting new data.

Due to the fact that the the Unknown Test Set cannot be verified, it is difficult to assess which model is more accurate. However, the Career Model yielded alarming results, mathematically. Consider, the overall Career dataset had a total of 76 jersey numbers that were actually retired between 1982 to 2018 out of a total of 972 players (7.8%). Therefore, it seems unlikely that 31 players out of 180 (17.2%) will get their jersey number retired in the future, according to the Career

Models predictions in the Unknown Test Set.

In comparison, the overall Team dataset contained 77 jersey numbers retired between 1982 to 2018 out 1423 (5.4%). The overall Team Model predicted 14 future players out of 228 (6.1%) retired, which is much more consistent. As a result, it appears that the Team Model is better than the Career Model, albeit they both had a prediction accuracy overall.

Limitations & Suggestions for Future Studies

As mentioned in the data section, methods were applied in an attempt to control for outliers. Training a model to consider outliers leads to overfitting. At any rate, a more robust method of controlling for outliers should be applied by future studies.

Only regular season statistics were measured in the data. Future studies that incorporate playoff statistics could have different the prediction accuracy than the models developed in this study. This study averaged out most of the statistical categories considered, it is unclear if the difference between a players regular season and playoff averages are statistically significant.

There are so many different predictive analytic methods that can be applied to the data used. Future studies that apply different methods may be more accurate than the logistic regression with BIC and forward selection implemented in this study.

Finally, The last year for the data used is 2017. Using more recent data would probably yield better predictions on the test set. Although this is a key limitation, the variable labeling whether a players jersey number is retired or not was obtained from

References

1. Basketball reference. (2019). *2018-19 NBA player stats: per game*. [Data set]. Retrieved May 2, 2021, from URL.

2. *Glossary*. (n.d.). Basketball reference. Retrieved April 28, 2021, from URL.

3. Goldstein, O. (2018). *NBA Players stats since 1950*. (Version 2). [Data set]. Retrieved March 28th, 2021, from kaggle.com/drgilermo/nba-players-stats.csv.

4. Hastie, T., Tibshirani, R., & Friedman, J. H. (2017). *The elements of statistical learning: Data mining, inference, and prediction*. (2nd ed.) New York: Springer.

5. Lambert, J., Gong, L., Elliott, C. F., Thompson, K., & Stromberg, A. (2018). rFSA: An R package for finding best subsets and interactions. *The R Journal, 10*(2), 295-308. ISSN 2073-4859

6. Land of basketball. (2021). *Players with the most selections for the NBA All-Star Game*. [Data set]. Retrieved April 20, 2021, from URL.

7. Mcleod, A.I., Xu, C., & Lai, Y. (2020). Package 'bestglm'. p. 1-48.

8. Mcleod, A.I., & Xu, C. (2010). bestglm: Best subset GLM. p. 1–39.

9. Miller, A.J. (1984). Selection of subsets of regression models. *J. R. Statist. Soc. A. 147*(3), 389-425.

10. NBA hoops online. (n.d.). *NBA number retirements*. [Data set]. Retrieved April 3, 2021, from URL.

11. *Retired Jersey Numbers* (n.d.). Real GM. URL.

12. Shar, T. (Dec. 6, 2017). *About train, validation and test sets in machine learning*. Towards data science. URL.

13. Weak side awareness. (June 16, 2011) *All NBA championships by players and coaches*. [Data set]. Retrieved April 21, 2021, from URL

14. Zhang, Z. (2016). Variable selection with stepwise and best subset approaches. *Ann Transl Med., 4*(7), 1-6.