Predicting the 5-year career longevity of the NBA Rookies

Anuj Parikh and Rahul Akkineni

University of New Haven

300 Boston Post Rd, West Haven, CT 06516

apari6@unh.newhaven.edu, rakki1@unh.newhaven.edu

Abstract

Selection of a sport team is a crucially important task of any team coach. For NBA team selection, historical data is one of the fundamental factors to make prediction about game results. For predicting an individual player, the historical data of player's performance is directly correlated to his performance in future matches. Past data can help predict the player's future performance using Data science techniques.

Introduction

The National Basketball Association (NBA), an American men's professional basketball league, is composed of 30 teams and is one of the four major professional leagues in United States and Canada. NBA affects the sports cultures in western countries. For selecting best players for a team based on their historical performances is one if the most crucial task of an NBA team coach.

A recent study demonstrates the importance of data in sports-related decision making of Rusty Simmons [1]. With use of Machine Learning Algorithms, we propose a tool for making good decision about team selection. This work devises a way to corelate historical data with the future performance of players. Furthermore, this work compares 3 algorithms with their parameter variations. Finally, this work reports experimental results and concludes with comments on past performances' implications on future performances.

Related Work

Rusty simmons [1]'s describes about the importance of data in decision making in the field of sports. He argues that sequential decision is also important to win a game. If the coach, who are the decision makers second guess themselves, then the decision might have negative implications. Data gives confidence to decision makers and increases likelihood of correct decision which dramatically affects game results.

Srdjan Lemez and joseph Baker [2] in their recent study predict the lifespan of an athletes. Where they perform an analysis on comparison between participation of elite level of sports and

mortality. Lemez study uses data of different sports' player based on sex, age and height. On Contradictory of traditional belief, Lemez research conclude that Participation in elite sports leads to longer lifespan.

Yuanhao (Stanley) Yang's [3] provides analysis of correlation between player's statistic and game performance. Yang uses collection of past 20 seasons of NBA from 2015. With this data yang's research got satisfactory result with predicting ongoing game result. Yang had PER (Player Efficiency Rating) which is measured per minute and have important relation with game's victory as an attribute. Yang tested on basketball-reference.com dataset. This research listed array of methods for predicting game and its failure reasons. It also depicts importance relation between PER and game performance.

Lemez et al. [4] provide greater insights into health of persons with above average height. As height is one of the key competencies of Basketball players. Lemez study further establish relation between height and mortality. Lemez used basketball-reference.com as a dataset. Lemez performs Kaplan-Maier cox regression survival analyses on Height statistical relations combines with gender and birthplace and analyze that. Lemez research\ provide a reason to investigate relation between height and mortality. While at the same time, research is conducted on very narrow category of individual, So It demands more investigating.

Andrew Powell-morse [5] provides rich comparison and analysis of various parameter of NBA player's physics. This includes Physical Stature, Origin, and Education. With having visual representation of data, Andrew provides vivid picture of NBA player's physical states.

Zafar's [6] predicted career length with ~70% accuracy. Experiments of Zafar uses Logistic regression with 19 features of NBA data to achieve 70% accuracy. After analysis of feature importance, we selected only 5 most relevant features which represent only 27% of real complete data. We successfully maintained 70% accuracy by using Naïve Bayes Algorithm with just 27% of actual NBA data.

Proposed method

In this paper, the necessary packages that were required were loaded and dataset was imported. Then, we checked the assumptions: the observations are independent, linearity of the independent variables and log odds. After that we did data exploration using OLS Regression to eliminate features having zero value. For prioritize features according to its importance in predicting target, we used KBest algorithms and then chose top features which play dominant role in desired forecast. Then we experimented with number of features i.e., k in KBest to figure out appropriate k for better optimization. Afterwards, we split the data into training and testing to determine the confusion matrix and harnessed a cross-validation with 5 splits.

Then, after data exploration we did data modeling where we tested fifteen different classification models including Logistic Regression, Naïve Bayes, XGBoost, Support Vector Machines, and many more with default parameters. After that we selected few methodologies which is having high accuracy for our optimization phase. In optimization phase, we used three

different optimization techniques: GridSearch CV, RandomSearch CV and Bayesian Optimization on three different methodologies: Logistic Regression, Naïve Bayes Gaussian, and Stochastic Gradient Descent. In that Logistic Regression and Naïve Bayes Gaussian showed more accuracy using RandomSearch CV.

Experimental results

We begin by selecting important features using L1-based feature selection and Tree-based feature selection models. Figure1 shows the importance of different features used in the model. As we can see that GP is the most important feature which is Game Played.

Features	Importance
GP	0.09583504
Min	0.05336147
PTS	0.05921336
FGM	0.05435793
FGA	0.04830798
FG%	0.06062864
3P Made	0.03063386
3PA	0.04315281
3P%	0.04378697
FTM	0.05529345
FTA	0.05039769
FT%	0.04992243
OREB	0.05490673
DREB	0.05167443
AST	0.05884954
STL	0.04842301
BLK	0.04497474
TOV	0.04562186
Const	0.05065805

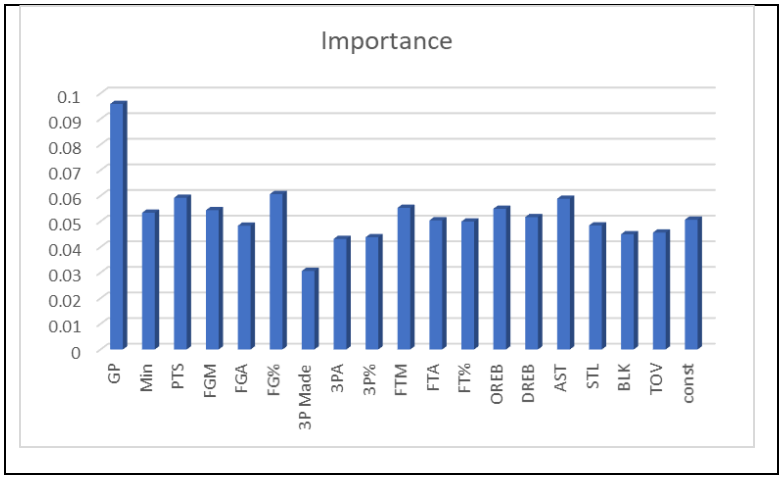

Figure 1 Feature Importance

After knowing the importance of different features, we implemented fifteen different modeling techniques on five features, ten features and all features separately Before standardization and After standardization. Where we implemented different modeling techniques with different parameters. Following hyperparameters are used in above mentioned classification methods:

In Logistic Regression, we used bfgs (Broyden–Fletcher–Goldfarb–Shanno) algorithm for nonlinear optimization. While setting maximum iteration up to 1000. In Naïve Bayes, we set hyperparameter of variance smoothing to 1e-9 for achieving stability. In SGD classifier, we used hinge loss based l2-norm penalty. We performed SGD with boundary of 1000 maximum iterations.

In KNN classification, we set deciding neighbours to 5 As odd number is easier to deal with when it comes to same votes for new classification object. In Decision Tree, Extra-Tree Classifier and Random Forest, we set decision tree's max depth unbounded, so we can leverage all the branches split data in decision tree as well as random forest. In SVM, we used radial basis kernel in Support Vector machine. For Bagging and Boosting Classifier, we set 10 and 100 for bagging and boosting algorithms' estimators, respectively. In AdaBoost, we used SAME.R algorithm for optimizing adaboost with 50 estimators. In LGBM classifier, we leverage lgbm gradient boosting classifier on 200000 bin-sample-size and 50 estimators with having upto 31 leaves for base learners. In Multilayer Perceptron Neural Networks, we used relu activation function with adam optimizer with batch-size set to 200. We set learning rate to 0.001.

Modeling Results: We get best accuracy of 70+ % using Logistic Regression on all available features, selected five features, and selected ten features from dataset. Figure 2 shows all fifteen classification method comparisons.

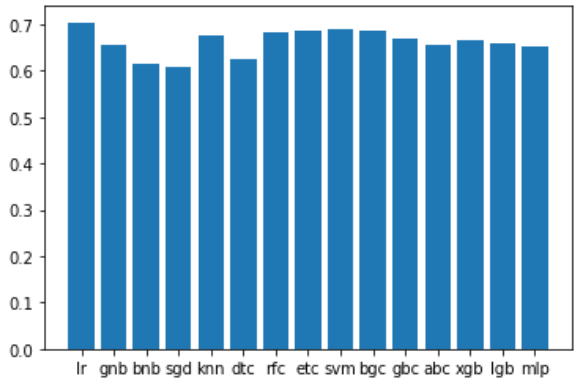

Figure A Before Standardization

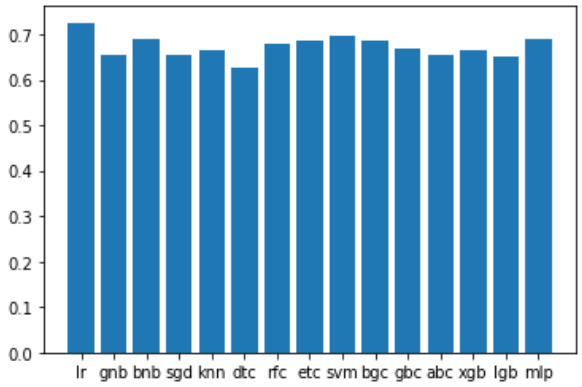

Figure A After Standardization

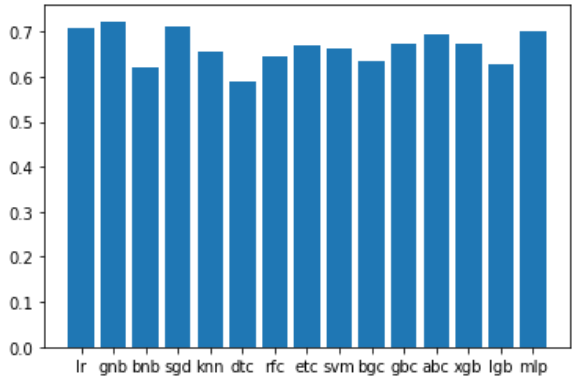

Figure B Before Standardization

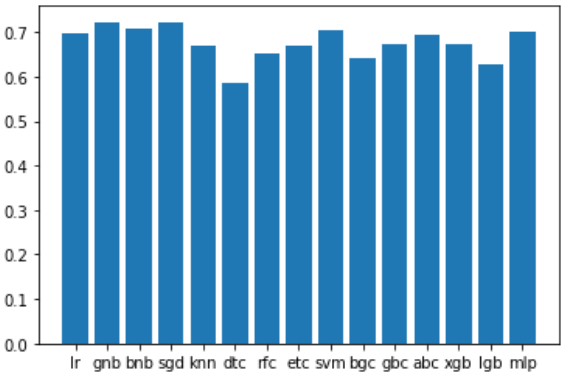

Figure B After Standardization

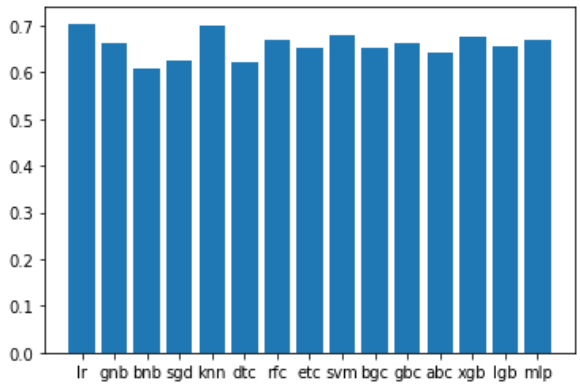

Figure C Before Standardization

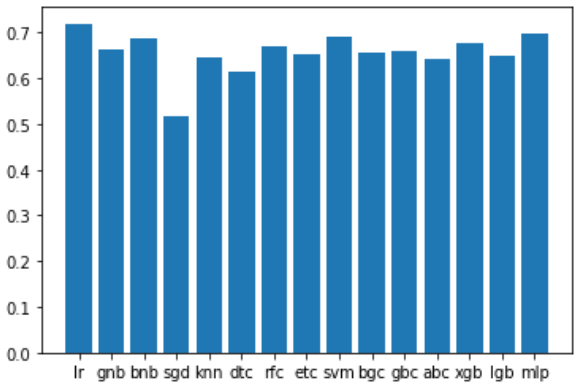

Figure C After Standardization

Figure 2: (A) Comparison of different modelling techniques on all features. (B) Comparison of different modelling techniques on five (5) features. (C) Comparison of different modelling techniques on ten (10) features.

After comparing fifteen different modelling techniques, we choose three different modelling technique which had the best accuracy and did Optimization on those selected models. We choose Logical Regression, Naïve Bayes Gaussian, and Stochastic Gradient Descent with specific parameters/Hyperparameters. In Logistic Regression, we used different algorithm for nonlinear optimization as a solver and other parameters such as 'C', 'max_iter' and 'Penalty'. For penalty 'l2' is the default one which is used by most of the solver with C parameter to control the penalty strength as a parameter. In Naïve Bayes Gaussian, we set hyperparameter of variance smoothing to 1e-9 for achieving stability as a parameter. In SGD (Stochastic Gradient Descent), we defined parameters such as 'alpha', 'n_iter_no_change', 'loss' and 'Penalty'. We performed SGD with boundary of 1000 maximum iterations as a parameter.

In optimization we used three different techniques: GridSearch CV, RandomSearch CV and Bayesian Optimization. GridSearch is used to modify supervised learning parameters and improve the efficiency. It tries all possible parameters of interest and find the best possible fit. In GridSearchCV implements we used a 'fit' method. For Logistic regression we used 'C' as a parameter. In Naïve Bayes we used 'Variance smoothing' to have a high accuracy. We used scikit learn to implement GridSearchCV. RandomSearch CV is like GridSearchCV but in contrast to GridSearchCV, not all parameter values are checked, but a fixed number of parameter settings are sampled from the distributions you specify. n_iter provides the number of parameter settings that are attempted. In Bayesian Optimization, the algorithm makes an informed guess for each run of hyperparameters on the objective function, which set of hyperparameters is most likely to increase the score and should be attempted in the next run. It is achieved by having regressor forecasts on several points and selecting the point based on the so-called acquisition function that is the best guess.

Optimization Results: We got an accuracy of 72.56% for our model After Standardization by using All features of the dataset which is highest in all modeling techniques we used. By taking

five best features, we got an accuracy of 72.25% using Naïve bayes gaussian modeling technique and with 10 features we got an accuracy of 71.95% using Logistic Regression. Figure 3 shows the results of Logistic Regression, Naïve Bayes Gaussian, and Stochastic Gradient Descent using all three optimization techniques and with selected features.

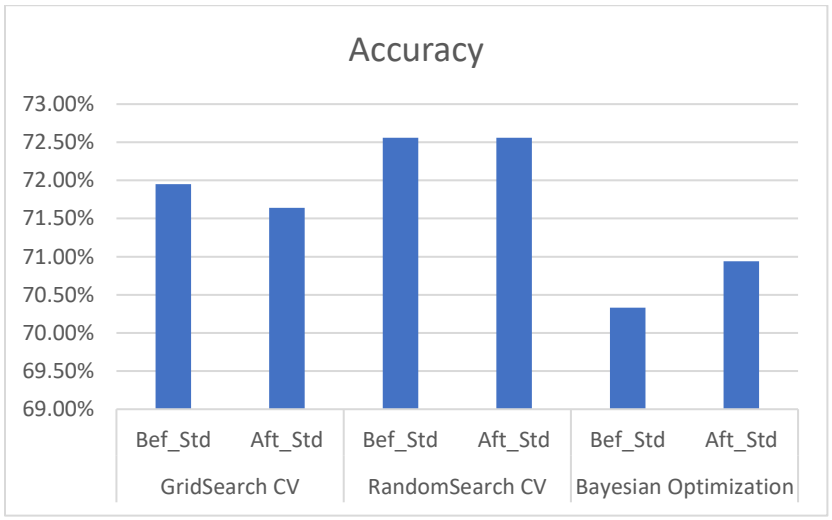

Figure A Optimization with All features

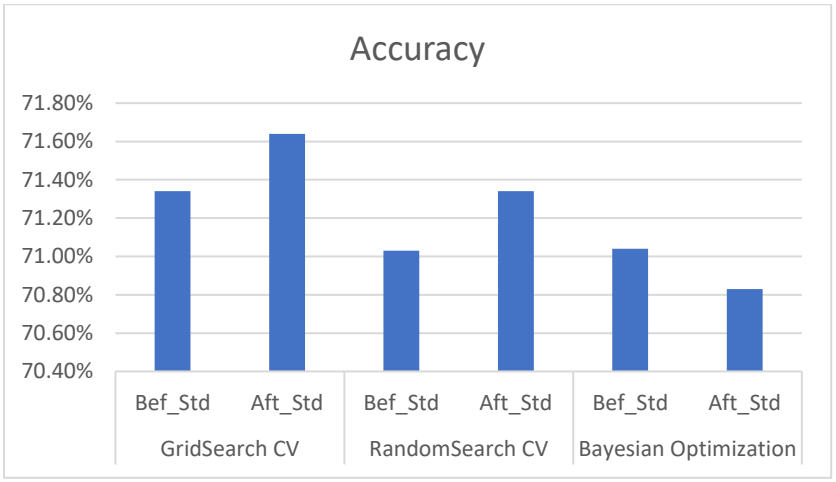

Figure B Optimization with selected five (5) different Features

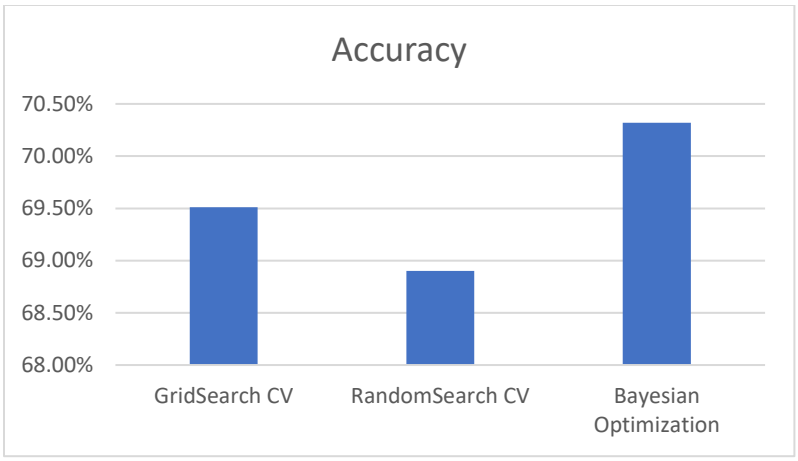

Figure C Optimization with five (5) features using SGD

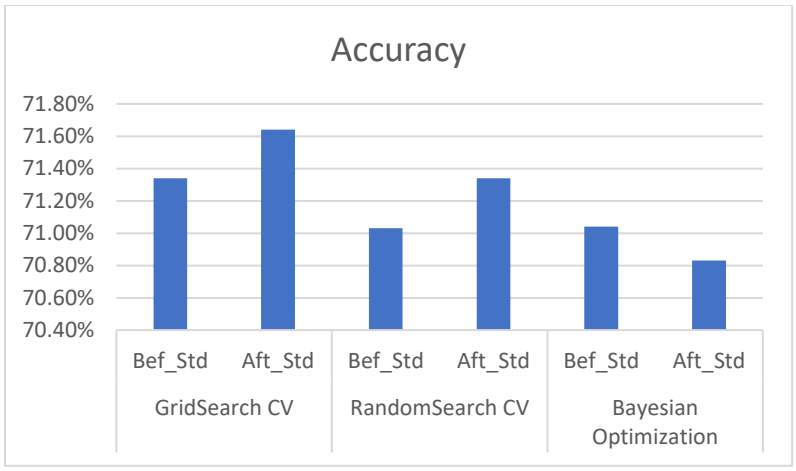

Figure D Optimization with ten (10) different features

Figure 3: (A) Optimization using Logistic Regression on All features. (B) Optimization using Naïve Bayes Gaussian on selected five (5) features. (C) Optimization using Stochastic Gradient Descent on selected five (5) features. (D) Optimization using Logistic Regression on selected ten (10) features.

Discussion

After calculating feature importance, It is clear that more number of game played by player give more accurate insights for prediction of longevity. It is clear from experimental analysis that Most important features for prediction are: Field Goals Percentage, Points per Game, Assists, and Free Throw Made. These features become more reliable when we have data of more games.

Although, some information has proven to be statistically less important. By adding additional 13 less statistically important features to the training dataset will generate only 0.31 % difference. This comparatively little percentage suggest that top features contribute to the overall performance of any model.

After testifying 15 different models with appropriate parameters, Logistic Regression equipped with all 18-feature dataset outperform its competitors with achieving **72.56 %**. Naïve Bayes perform exceptionally well with having only 5-statistically-important-feature dataset with achieving **72.25 %** accuracy in classification.

Analysis of various combination of data-features and 15 Machine Learning model provide a string argument for using features such as Field Goals Percentage, Points per Game, Assists, and Free Throw Made are crucial for game performance prediction. We can improve predictions with collecting data of more games.

Conclusion and future work

Initially we performed various tasks like Data exploration, data modeling and data optimization. After performing various modeling techniques, we get 72.56% accuracy on learning methods, which strongly support the argument of relating player historical performance to longevity of player. Since we observed accuracy of 70+%, we also take few more performance metrics such as precision, recall and f-score under consideration which help us to ensure the result of our outcomes. However, 72.56% is pretty good to deduce that pattern exist in asked research question. We strongly believe that Data Model with larger data can more accurately pinpoint relations between performance and longevity.

Overall, we are satisfied with our results. Future works latent includes expanding the parameters and dataset studies. Other way to improve model would be to find more observations or by reducing variables and reducing overfitting. We also think that it would be interesting to model using 3 features or else predict the longevity for next 2-years or for next10-years.

Appendix for link to the GitHub repository

DAT Team, Predicting the 5-year career longevity of the NBA Rookies, (2020), GitHub Repository, https://github.com/Anuj-parikh/DAT

Reference

1. SIMMONS, RUSTY: “Golden State Warriors at the Forefront of NBA Data Analysis.” http://www.sfgate.com/warriors/article/Golden-State-Warriors-at-the-forefront-of-NBA5753776.php

2. Lemez, S., Baker, J. Do Elite Athletes Live Longer? A Systematic Review of Mortality and Longevity in Elite Athletes. *Sports Med - Open* **1,** 16 (2015). https://doi.org/10.1186/s40798-015-0024-x

3. Yuanhao (Stanley) Yang: Predicting Regular Season Results of NBA Teams Based on Regression Analysis of Common Basketball Statistics. https://www.stat.berkeley.edu/~aldous/Research/Ugrad/Stanley_Yang%20_Thesis.pdf

4. Lemez S, Wattie N, Baker J (2017) Do "big guys" really die younger? An examination of height and lifespan in former professional basketball players. PLoS ONE 12(10): e0185617. https://doi.org/10.1371/journal.pone.0185617

5. Andrew Powell-Morse: “The historical profile of the NBA player: 1947-2015”

6. https://syedzafarcom.wordpress.com/2017/09/06/ml-classification-predicting-5-year-career-longevity-for-nba-rookies/