Projecting In-Prime Average Annual Value (AAV) of NBA Juniors To Negotiate Rookie Extensions

Krish Khanna
University of Chicago Laboratory High School

Note that the quality of this paper is limited by the lack of a complete high school education of the sole author.

Abstract

The new Collective Bargaining Agreement in the NBA allows rookie contract extensions to be negotiated between the third and fourth years below the previous regulated floor amount of 25% of the salary cap. This creates a need to project In-prime performance of players at the conclusion of their third contract year.[1] In this paper, I scraped NBA performance data for the first three years for all players from 1996 to 2019, and I scraped their inflation adjusted salaries between the ages of 26 and 29 yrs. I used visual analysis, statistical techniques, and advanced predictive modeling to identify factors that contribute to player performance in their prime, and then explored models capable of projecting player values in their prime with suitably low error. The best regression model achieved a mean absolute error of approximately five million dollars.

Introduction

In the NBA, rookie contracts are controlled by a team for a period of four years, two of which are guaranteed and two of which are rookie scale team options with no player leverage. Standard NBA rookie contracts are structured such that they can either be extended during a brief window of time between their third and fourth year, or the contract expires and the rookie becomes a Restricted Free Agent at the completion of the fourth year.[2] This puts massive value at risk for the original team: Signing underperforming rookies to extensions too early can be damaging to a team's finances and leave them without enough cap space to sign other players, while failing to extend superstars in a timely fashion allows other teams to compete for their signature, driving the star athlete's price through the roof.

[1] "Contract Types," CBA Breakdown, accessed February 4, 2024, https://cbabreakdown.com/contract-types/#extensions.
[2] "Contract Types," CBA Breakdown.

The extension is also attached to a number. For any team considering the path of rookie extensions, that number needs to be right. Under the old Collective Bargaining Agreement (CBA) between the NBA and the Basketball Players Association, teams had only two options when it came to rookie-scale extensions. They could offer as many as four years with a starting salary up to 25% of the salary cap, or they could offer a five-year deal that had to start at 25% of the salary cap.[3] If they wanted to offer that fifth year, they had no flexibility in negotiating the starting salary. The new CBA, however, grants teams the ability to offer a five-year deal that doesn't necessarily begin at 25% of that year's salary cap.[4] This puts the onus on the original team to get the number just right and to negotiate effectively with the player in order to maximize the risk-reward ratio of an early rookie extension.

Related Work

Barring private use models from individual NBA teams, there is no public domain research focused on prime AAV of players. Most prior research focuses on ‘fair market value’ of players in the contract year based primarily on their NBA stats or on the projected value of a prospect based on Pre-NBA statistics and Athletic Measurables, such as vertical jump height and wingspan. These studies are focused on a very narrow set of players with a limited timeframe.

One such study attempted to predict the salary of NBA players based upon their game statistics. It used 2018/19 game and salary data. The data was cleaned, players with 5 or less games played were dropped on account of excess variance, position data was refactored, and metrics were converted to stats per minute for ease of comparison. The author tested Linear Regression, Support Vector Regression, AdaBoost, Random Forest and XGBoost models. Isolation forests were used to identify and jettison outliers. Models were benchmarked against a mean prediction dummy regressor to evaluate performance. All regression analysis outperformed the baseline Mean Predicting Dummy Regressor model’s test set evaluation metrics (RMSE: 1.57, R Squared: -.01). The top performing individual model was the AdaBoostRegressor (RMSE: .99, R Squared: .60), while the top performing overall model was an AdaBoost/Polynomial SVR/Linear SVR Voting Regressor (RMSE: .95, R Squared: .63). Minutes/game, games played, and age best explained the dependent variable. A voting model with adaboost, polynomial SVR, and linear SVR components performed the best.

---

[3] Bryan Toporek, "The NBA's New Labor Agreement Paved the Way for a Record Rookie-Scale Extension Deadline," Forbes, last modified October 24, 2023, accessed February 4, 2024, https://www.forbes.com/sites/bryantoporek/2023/10/24/the-nbas-new-labor-agreement-paved-the-way-for-a-record-rookie-scale-extension-deadline/?sh=4730c4a751a0.

[4] "Contract Types," CBA Breakdown.

Data Set Creation

I used a Kaggle dataset that contains every NBA draft pick from 1989 to 2021 as the master list of players that I used to guide my automatic scraping program.[5] This list had 1,923 player names. For the dependent variable, I used NBA salary in the player's prime. Since there was no readily available dataset for this, I built a custom web scraper using BeautifulSoup library to scrape salary data from hoopshype.com.[6] This website has actual salaries as well as inflation adjusted salary data. The inflation adjustment is done using US Department of Labor inflation statistics. For each player, I scraped their inflation adjusted salary for ages 26, 27, 28, and 29 years and added this to the DataFrame, as prior research has shown these years to be the prime of an NBA player's career.[7] Given the change in rules relating to salary caps, I explored whether to use salary as a percent of the salary cap in that year. However, while maximum inflation adjusted salary has grown from year to year, the mean has stayed about constant (fig 3). "In Prime AAV" was calculated by averaging the inflation adjusted salary over four years. If the player is younger than 29 years, then I calculated the average salary based on numbers of years played in their prime, but for older players, I included '0' salary values in any given year and calculated their mean salary over all four years.

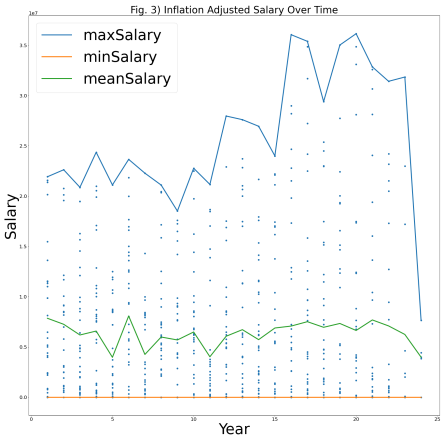


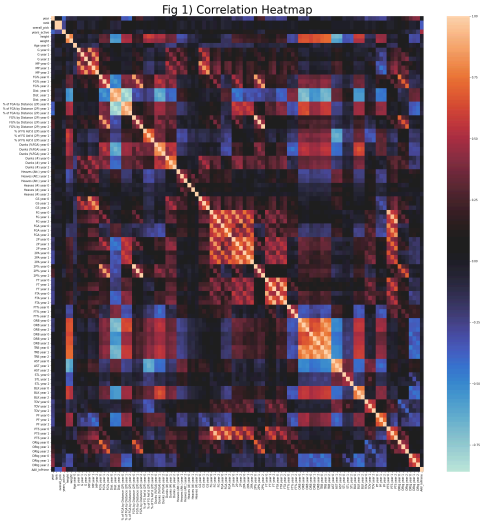
Fig. 3) Inflation Adjusted Salary Over Time

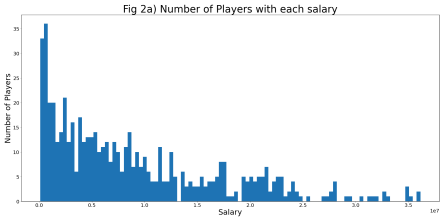
Next, I scraped NBA rookie player statistics from basketballreference.com. I normalized this NBA data, including box score data and shooting data, to 100 possessions for each of the three years for a rookie, and included their height and weight. I flattened this data, and named each variable using the nomenclature - statistic_corresponding year. I appended these independent variables to the dataframe. Then, I eliminated all players (rows) using the following two rules - 1) Players before 1996 since their data was either missing or incompatible, 2) Players who were missing NBA statistics for any of the three years. I then removed columns that were duplicative or redundant, such as age in years 1 and 2 and league name. I normalized the year by subtracting 1995 from each year to remove unnecessary noise. At the end of this process, I had 806 players, 100 independent variables, and the response variable (mean salary during prime).

[5] "NBA Draft Basketball Player Data 1989-2021," Kaggle.com, accessed February 4, 2024, https://www.kaggle.com/datasets/mattop/nba-draft-basketball-player-data-19892021.
[6] "NBA Player Salaries," HoopsHype.com, accessed March 3, 2024, https://hoopshype.com/salaries/players/2016-2017/.
[7] Andy Feng, "Peak Age in Sports," Dartmouth Sports Analytics, last modified November 10, 2021, accessed March 3, 2024, https://sites.dartmouth.edu/sportsanalytics/2021/11/10/peak-age-in-sports/.

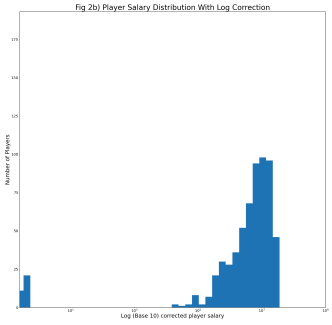
Exploratory Data Analysis

After compiling the DataFrame, I used preliminary data exploration techniques to understand the data distribution. First, I created a heatmap to examine potential multicollinearities (Fig 1). The paneled grid is indicative of a high correlation between NBA statistics in the first, second, and third years. This suggests that we can focus on 3 by 3 grid square correlations. The heatmap also shows that Height and Weight are positively correlated to each other, and are positively correlated with rebounds, but negatively correlated with average field goal distance. This makes sense since guards take longer shots and operate farther from the basket. In addition, we see that field goal percentage (FG%) is inversely correlated with FG distance, so some interaction of the two stats will likely determine the players scoring value. An interesting correlation is the negative correlation between Defensive Rating and rebounds, as traditionally centers, who get the most rebounds, are defensive anchors. This correlation is largely unimportant for the current task, but can be explored further in a future project.





Next, I plotted the mean annual salary in a histogram to examine the skews. The salary data was incredibly right skewed (Fig 2A). I applied a log transformation to the salaries to project the data onto a more gaussian distribution, which generated two clear clusters (Fig 2B). The first cluster is players that did not get an extension contract, and therefore did not play in the NBA during their prime years. The other cluster is normally distributed with a mean In-Prime AAV of $10 million. Given these clusters, I decided to create two distinct models. The first model will aim to classify players into those who are expected to remain in the league versus those who will not stay in the league in their prime years. The second model will predict the In-prime AAV of those players who are predicted to continue in the league in their prime.

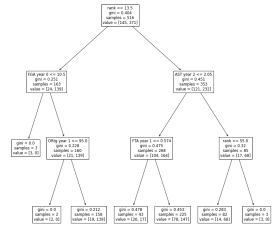


Results

Model 1: Predict whether the player will be in the league in their prime

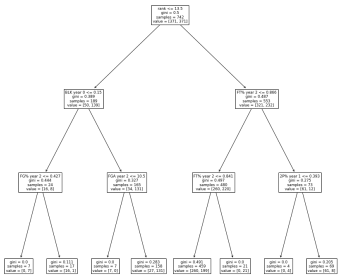
First, I fit a decision tree to the data to explore which traits were the most important in predicting whether the player stayed in the league (Fig 4). Given the right skewed data set, the initial decision tree tended to predict that all players would remain in the league with acceptable accuracy, which defeats the purpose of the model.

Fig 4) Decision Tree without Upsampled Data



In order to develop a model that actually attempted to learn a valuable decision boundary, I upsampled the data with players no longer in the league to even out the training set. The resulting decision tree classifier was more balanced after upsampling (Fig 5).

Fig 5) Decision Tree with Upsampled Data



After upsampling, I ran a variety of models with different hyperparameters, including logistic regression, bagging, boosting, decision trees, and support vector classifiers. Of these models, support vector machines performed the worst. The baseline accuracy per the first reference paper cited above was 60%. The logistic regression model scored 63.5% accuracy on both the validation and test set. The best bagging and boosting models achieved close to 68% accuracy (Fig 6, 7, 8).

Fig 6) Logistic Regression

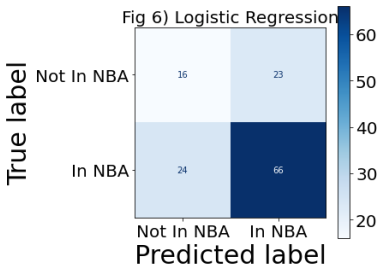


Fig 7) Bagging

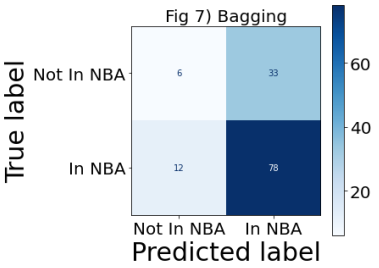


Fig 8) Boosting Model

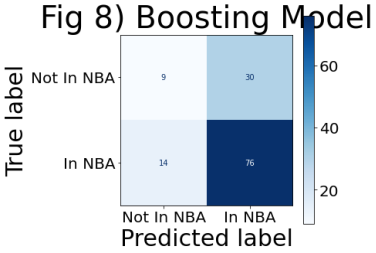
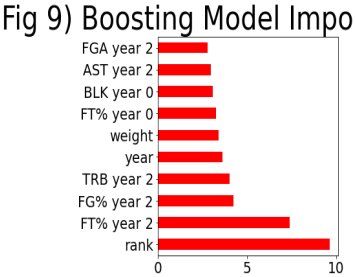


Fig 9) Boosting Model Important Traits



Both these models had similar accuracy and for players that they misclassified, both models tend to predict them as remaining in the league rather than not. This is a more conservative outcome and

beneficial to the next model since regression can assign low values to these players in the next model. The four traits that contribute most to the best boosting model are draft rank, free throw percentage, field goal percentage, and total rebounds (Figure 9).

Model 2: Predict In-Prime AAV for players predicted to be in the league

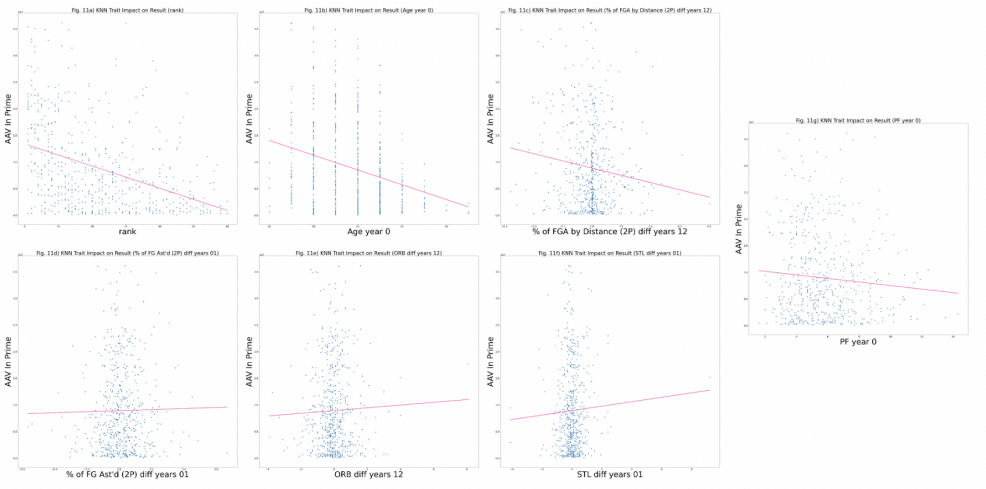
I built a linear regression model with all traits to set a floor for evaluation of further models. This model produced an RMSE of just over $8.1 million. I then used a lasso regression model to identify traits that do not meaningfully contribute to predicting In-prime AAV. Given the high collinearity between traits in each of the first three years, if a trait was not zeroed out by the lasso regularization, all years with that trait were used in future steps. After elimination, the traits considered significant were draft rank, age, field goal percentage, percentage of shots attempted that were two pointers, percentage of two pointers made, percent of two pointers that were assisted, number of dunks, 2 point field goal percentage, and defensive rating.

rank	-1.968658e+06
Age year 0	-1.301067e+06
FG% year 2	-7.112407e+04
% of FGA by Distance (2P) year 2	-6.753870e+05
FG% by Distance (2P) year 0	3.205592e+05
% of FG Ast'd (2P) year 1	4.576777e+05
Dunks (#) year 0	2.490399e+05
Heaves (Att.) year 2	-2.565077e+04
Heaves (#) year 0	1.148397e+05
GS year 1	1.059660e+05
2P% year 0	5.462565e+03
DRtg year 0	1.691537e+05

Next, I tested K Nearest Neighbors Regressors, Gradient Boosting Regressors, Neural Networks, and Linear Regressions, but was not able to improve beyond an RMSE of 7.5 million dollars. I reconfigured the metrics data to include year 0 values, and change from year 0 to year 1, and from year 1 to year 2, instead of using year 1 and year 2 absolute values. This allowed the model to consider growth over the three years rather than standalone statistical values that are more difficult to contextualize. With this new dataset, only traits where one of the years was deemed statistically significant were included ($p<0.05$). I tested K Nearest Neighbors Regression, Support Vector Regression, Lasso Regression, AdaBoost Regression, and Neural Network models. The best performance was from K Nearest Neighbours Regressor. The traits used for 50 Nearest Neighbours Regressor were draft rank, rookie age, increase in number of two point field goals attempted from year two to year 3, increase in number of assisted 2 pointers from year 1 to year 2, increase in steals from the rookie year to year 2, the number of personal fouls per 100 possessions in the rookie year, and the improvement from the second year to the third year of offensive rebounds. The model resulted in validation RMSE scores of approximately 7 million, and had a test mean absolute error of 4.98 million dollars. Top quartile In-Prime AAV player salaries range from 10 million dollars to approximately 36 million dollars, and the mean In-Prime AAV for the dataset is $10 million. Therefore, an absolute mean error of 5 million dollars on the testing set is valuable in setting a price ceiling for player negotiations.

Discussion

While draft rank and rookie age were already established as good predictors by previous studies, the regression models used in this study established other important traits that are likely significant predictors of NBA success. These other factors include increase in number of two point field goals attempted from year two to year three, increase in number of assisted two pointers from year one to year two, increase in steals from the rookie year to year two, the number of personal fouls per 100 possessions in the rookie year, and the improvement from the second year to the third year of offensive rebounds (Fig 11 below).



It makes logical sense that higher draft picks are less likely to have high contracts, but it is interesting that players that were drafted at a younger age tend to earn more in their prime. Perhaps it is the case that more time in NBA caliber facilities significantly improves the player salary outcomes, as they have adequate time to adapt to the NBA and build on their skillset. It is interesting that fouls are negatively correlated to In-Prime AAV because fouls can be used as a proxy for effort. But, this finding likely suggests that not fouling out of games is a valuable skill at the NBA level. Ultimately, the optimal classification model achieved 70% test accuracy for predicting whether a player would stay in the league for their prime, and the optimal regression model achieved an MAE of less than 5 million dollars. Both of these models are suitably accurate to be used, and represent baseline accuracy values to be improved upon in future improvements to this modeling task.

Acknowledgements

This paper is an edited version of a work written for MPCS 53120 in 2024 at the University of Chicago. I thank Davender Singh Sahota PhD for his guidance on all aspects of this project.

References

"Contract Types." CBA Breakdown. Accessed February 4, 2024. https://cbabreakdown.com/contract-types/#extensions.

Diallo, Mamadou. "NBA Rookie Longevity Project." Kaggle.com. Accessed March 3, 2024. https://www.kaggle.com/code/mamadoudiallo/nba-rookie-longevity-project#3.-Exploratory-Data-Analysis.

Feng, Andy. "Peak Age in Sports." Dartmouth Sports Analytics. Last modified November 10, 2021. Accessed March 3, 2024. https://sites.dartmouth.edu/sportsanalytics/2021/11/10/peak-age-in-sports/.

"NBA Draft Basketball Player Data 1989-2021." Kaggle.com. Accessed February 4, 2024. https://www.kaggle.com/datasets/mattop/nba-draft-basketball-player-data-19892021.

"NBA Player Salaries." HoopsHype.com. Accessed March 3, 2024. https://hoopshype.com/salaries/players/2016-2017/.

Pincus, Eric. "Grading NBA's Latest Rookie-Scale Extensions." Bleacher Report. Last modified October 27, 2023. Accessed February 4, 2024. https://bleacherreport.com/articles/10094004-grading-nbas-latest-rookie-scale-extensions.

Quinn, Sam. "NBA Rookie Extensions." CBS Sports. Last modified October 13, 2023. Accessed February 4, 2024. https://www.cbssports.com/nba/news/nba-rookie-extensions-how-much-should-the-best-players-from-the-2020-rookie-class-get/#:~:text=Teams%20have%20grown%20increasingly%20willing,total%20of%2032%20rookie%20extensions.

Sports Reference College Basketball. Accessed February 4, 2024. https://www.basketball-reference.com/players/d/doumbse01.html.

Toporek, Bryan. "The NBA's New Labor Agreement Paved the Way for a Record Rookie-Scale Extension Deadline." Forbes. Last modified October 24, 2023. Accessed February 4, 2024. https://www.forbes.com/sites/bryantoporek/2023/10/24/the-nbas-new-labor-agreement-paved-the-way-for-a-record-rookie-scale-extension-deadline/?sh=4730c4a751a0.