Krish Khanna
MPCS 53120 Applied Data Analysis
January 31, 2024

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

## Context

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 (RFA) at the completion of the fourth year.[1] 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.

The extension is also attached to a number. And 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 begin at 25% of the salary cap.[2] 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.[3] 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.

## Proposed Solution and Impact

In recent years, we've seen the prevalence of rookie extensions increase significantly from an average of 6.8 extensions per year between 2013 to 2019, to an average of 10.8 extensions per year from 2020 to 2022, and 13 extensions in 2023.[4] Taking a bet on the upside of these young players has increasingly become the norm because teams feel that the upside of locking in a young player on the way up is greater than the downside of that player plateauing or leaving for nothing. However, some experts argue that this trend has led to teams overpaying in the extension window out of fear of losing players despite a lack of production to back it up. For example, Jaden McDaniels, a defense first prospect on the Timberwolves averaging only 12 points per game, received a contract with an average annual value of 27 million dollars over the next five years. If McDaniels develops as Minnesota hopes, this could be seen as market value, but, locking up this much capital in a lanky 3 and D wing seems to some experts a tad overzealous given his current production.[5]

---

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

2 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.

3 "Contract Types," CBA Breakdown.

4 Sam Quinn, "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.

5 Eric Pincus, "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.

In this project, I aim to predict the average annual salary of a player over a two year period in their prime (27-28 yrs) using over 50 variables spanning - Combine data, such as body fat percentage, lane agility, bench press, sprint speed, wingspan, vertical jump; NBA statistics during the 3-yr rookie contract; College Performance Statistics, if they played in college; and in the case of international players drafted into the NBA, performance data from the International team they played with prior to signing the NBA rookie contract.

This predicted average annual salary will serve as an objective metric in guiding the original team whether the player is valuable enough to extend. In addition, the average annual salary would serve as a hard price ceiling in negotiating rookie extension contracts such that teams do not overpay in the extension window.

**Data Sets/Scrapers, Software, Analysis**

Given the team oriented nature of basketball, it is difficult to reliably quantify the individual impact of a player coming into the NBA with a narrow data set (such as using only their NBA stats or just Combine stats or just College Performance Stats), especially since many players choose to abstain from physical testing at the combine, and there is wide variance of competitions between different divisions of college basketball, the G league, and international leagues making it difficult to compare players side by side. In my project, I will try to scrape and use independent variables from multiple contexts (International team, College Team, NBA performance and Combine data) over a broad time frame (10 year data from 2010 to 2020) for each player to normalize factors outside the control of individual players.
I have identified the following data sets, many of which are not collated. I have indicated below the scrapers I will need to write to collate much of the data below:

- ***Average Annual Salary in the prime years*** (Dependent Variable): I will build a web scraper for the ESPN salary data websites and extract data dating back to 2010.[6]

- ***NBA Combine Data*** (Independent Variable): Another web scraper will be built to extract anthropomorphic and strength and agility stats from the NBA combine for each player.[7] Special care will need to be taken to ensure the missing values are dealt with in a way that doesn't influence the prediction quality, as it is not uncommon for top players to abstain from the combine.

- ***College Performance Data*** (Independent Variable): I will also build a web scraper for basketball reference college data for all players.[8]

- ***Draft Data*** (List of all players to consider): Draft picks dating from 1989 to 2021 are available in csv format on Kaggle.[9]

- ***NBA Performance Data*** (Independent Variables):
    - **NBA Stats (e.g.,** points per game, games played, minutes, rebounds, assists, true shooting percentage, win shares, and value over replacement player)**:** I will scrape data from college basketball reference and basketball reference to procure pre-NBA data both overseas and in college.[10]

---

[6] ESPN, accessed February 4, 2024, https://www.espn.com/nba/salaries/_/year/2013. (example salary url)

[7] https://www.nba.com/stats/draft/combine-anthro?

[8] Sports Reference College Basketball, accessed February 4, 2024, https://www.sports-reference.com/cbb/players/carmelo-anthony-1.html. (example player page)

[9] "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.

[10] Sports Reference College Basketball, accessed February 4, 2024, https://www.basketball-reference.com/players/d/doumbse01.html; Sports Reference Basketball, accessed February 4, 2024, https://www.basketball-reference.com/international/players/deni-avdija-1.html

- **RAPTOR data**: High level RAPTOR data which serves as a quantification of player impact pioneered by fivethirtyeight.com is publicly available in csv format on GitHub.[11]
- **Effort Stats**: Effort statistics such as distance run per game and other more nuanced statistics can be downloaded directly from the NBA website.[12]

*Software*: For this project, I will take advantage of the python libraries Pandas for its data frame features, NumPy, scikit learn to efficiently implement and fit models to the data, statsmodel for data statistic generation, matplotlib.pyplot and seaborn for visualization generation, and beautiful soup for web scraping. In addition to these packages, I will need to write custom web scrapers with beautifulsoup from scratch to extract some of the data indicated above.

*Analysis:* I plan to implement a wide variety of regression models to best formulate the predictions for AAV for rookie contract players in their prime years (age 27-28) and to understand which statistics best explain prime success.[13] These include linear regression, decision tree regression, random forest regression, artificial neural networks, support vector regression, XGBoost regression, and AdaBoost regression.

**Literature Review**

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.

***#1) NBA Salary Regression Modeling*** [14]

*Description, Data, Preprocessing*: This study built a regression model that could predict the salary of NBA players based upon their game statistics. It used 2018/19 game data from basketball reference and 2018/19 salary data from hoopshype. 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.

*Models/ Analysis:* Linear Regression, Support Vector Regression, AdaBoost, Random Forest and XGBoost models were tested; Isolation forests were used to identify and jettison outliers.

*Evaluation and conclusions*: Models were benchmarked against a mean prediction dummy regressor to evaluate performance. All of their 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.

*Implications for me*: There is an inherent challenge in untangling the collinearity between variables, one example being that players that play more tend to turn the ball over more, and I will need to find a way to

---

[11] https://github.com/fivethirtyeight/data/blob/master/nba-raptor/historical_RAPTOR_by_player.csv; Nate Silver, "How Our RAPTOR Metric Works," FiveThirtyEight.com, last modified October 10, 2019, accessed February 4, 2024, https://fivethirtyeight.com/features/how-our-raptor-metric-works/.

[12] https://www.nba.com/stats/players/speed-distance

[13] https://sites.dartmouth.edu/sportsanalytics/2021/11/10/peak-age-in-sports/

[14] https://medium.com/@blant.jesse/nba-salary-regression-modeling-4846e53a1d3b

resolve this. Based on this study, I will think about potentially using a voting model for final predictions. I will benchmark against mean as a baseline, but also examine whether any model exhibits marked improvement over a pure minutes linear regression model.

***#2) Logistic Regression Analysis- 5-Year NBA Rookie Classification*** [15]

*Description, Data, Preprocessing*: This study analyzed the impact of 21 variables, all pertaining to NBA performance, in predicting whether a player will last 5 years after coming into the league. It used a pre-collated list of NBA players with rookie data and whether they played 5 or more years in the league. The NaN/missing values in the dataframe were replaced with the median as it was resistant to outliers, which may influence the results, and a 70-30 train test split was used to divide up the dataset.

*Models/ Analysis*: Logistic regression was the only model used, however the author took care in selecting subsets for the model to be trained on.

*Evaluation and conclusions*: AUC and AIC were used to determine model quality and confusion matrices and f1 scores were used for model comparison. Games, points per game, and minutes played were the best standalone indicators followed by free throws made and field goals made.

*Implications for me*: I will need to think carefully about how to deal with NaN/missing values, and will potentially store international and college data in different columns and fill international data with 0 for college players and vice versa if the players didn't play in the other respective seasons to ensure that I'm not influencing the data in any way.

**Progress, Plan of Action, Perceived Challenges**

I have already researched and finalized data sources for all dependent and independent variables. Most of the independent variables do not have pre-collated data sets, and I will need to write multiple scrapers to collate data sets for college performance, NBA stats, and International Team performance. I have already written a scraper for NBA performance data. Next, I will be writing scrapers to collate player level data for college performance, international team performance data, Annual Salaries. I may need to look for additional data sources at that point if I have insufficient data upon scraping from the currently identified sources. Then, I will write code in python to combine the many different datasets using the player name as the join column. I will need to keep the data for international performance stats, NBA stats, and college performance stats separate. As a next step, I will start out doing basic correlations looking for collinearity and histograms to look at data distribution, outliers, etc. I will need to think through how to deal with collinearity, and will cleanse and preprocess the data based on these analyses. Thereafter, I will whittle down the number of useful statistics using various statistical methods, and write processes for feature selection and hyper parameter tuning alongside the standard model creation, training, and testing pipeline. I will then move on to iteratively fitting models, comparing the results, and fine tuning as needed. In the final phase, I will evaluate the model's performance as noted below.

**Evaluation**

I plan to evaluate my predictions against a mean salary prediction model, and a pure minutes played linear regression model using root mean squared error to quantify the improvement in regression quality, as there is no publicly available model to benchmark against. I will also plan to use 5-fold cross validation to benchmark models before running them on the test data for the most accurate error readings.

---

[15] https://medium.com/@ammasanaswar/logistic-regression-analysis-5-year-nba-rookie-classification-ab6d43c2a1e4