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

Krish Khanna

MPCS 53120 Applied Data Analysis

March 3, 2024

Context

- NBA rookie contracts structured as 2+2, window to extend for 5 yrs in 3rd yr, else RFA* at expiration
- Extension amount open to negotiations per new CBA**
- Rolling the dice on extensions is becoming the norm (avg 6.8 extns from ‘13 to ‘19 vs. 13 extns this year)
- Massive Value at risk for NBA teams:
  - Sign underperforming rookies too early = finances at risk, insufficient cap for other players
  - Failure to sign superstars in time = Compete with other teams, drive up prices
  - Salary negotiated critical to maximizing risk to reward ratio

* Restricted Free Agent
** Collective Bargaining Agreement between NBA and Basketball Players Association

Literature Review

- Only private use AAV in prime models

- #1: NBA Salary Regression modeling using 2018/19 NBA game data;
  - 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

- #2: Logistic Regression for 5yr NBA rookie classification using 21 NBA stats variables
  - Games, points per game, and minutes played were the best standalone indicators followed by free throws made and field goals made.

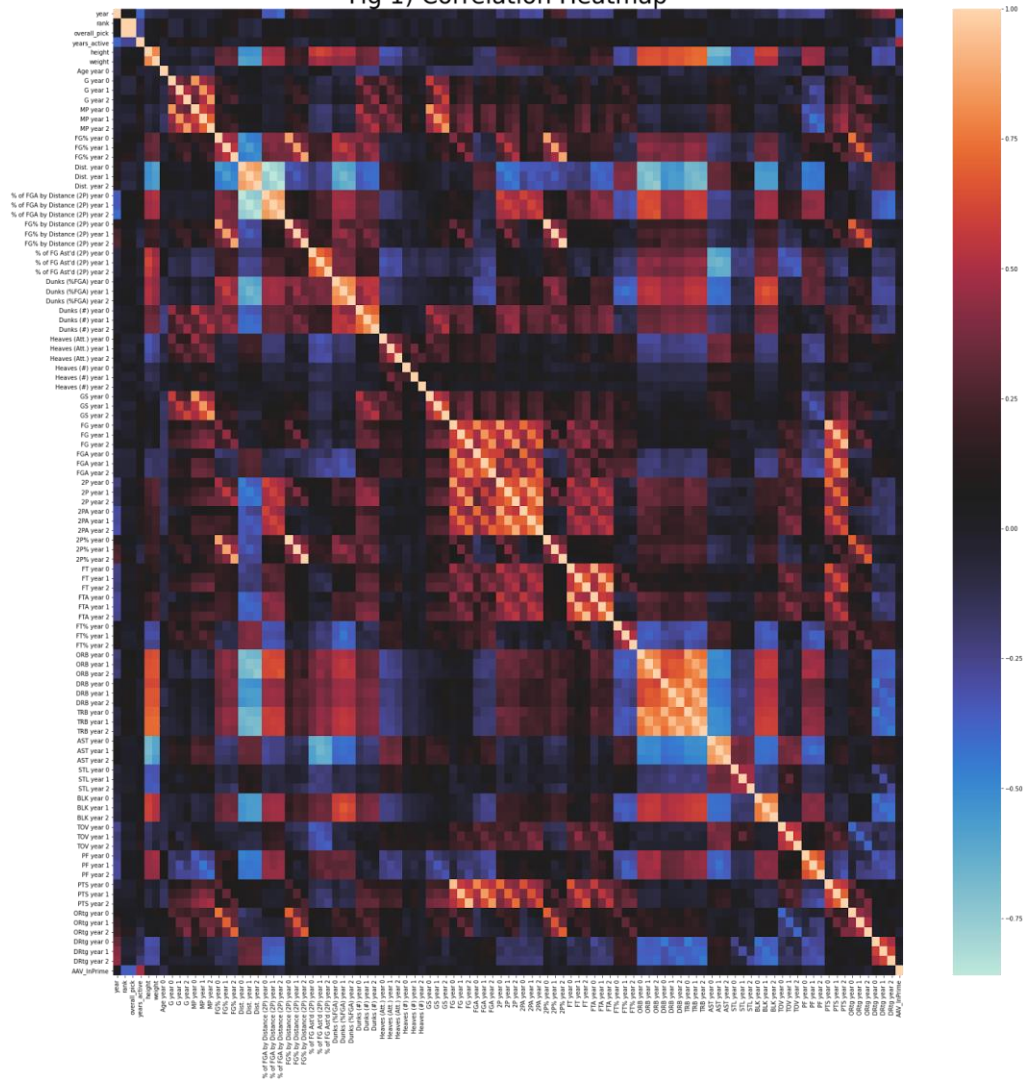
Data Set Creation

- Kaggle dataset for player names; NBA draft picks from ‘89 to ’21 (1,923 players)
- Built custom web crawler using BeautifulSoup library to scrape:
  - Inflation Adjusted Salary data from hoopshype.com
  - NBA statistics from basketballreference.com
- Calculated inflation adjusted salary during prime (26-29 yrs)
  - Max has grown from yr to yr; Mean has stayed about constant
- Normalized NBA metrics to 100 possessions
- Flattened the data set
- Removed players:
  - Before 1996 due to incompatible data
  - Players missing NBA stats for any of the three yrs
- Removed columns that were duplicative or redundant, e.g., age, league name

- Final DataFrame = 806 players, 100 independent variables, response variable
- Scraping was very time consuming; query limits placed on website; numerous edge cases
- College and International pro-league data was unclean

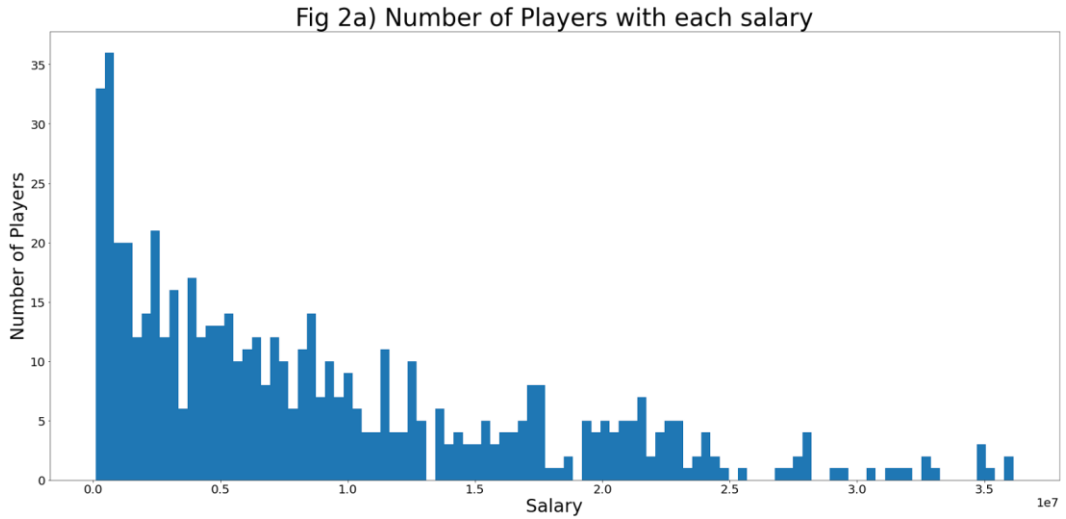
Exploratory Data Analysis: Heatmap (Collinearities)

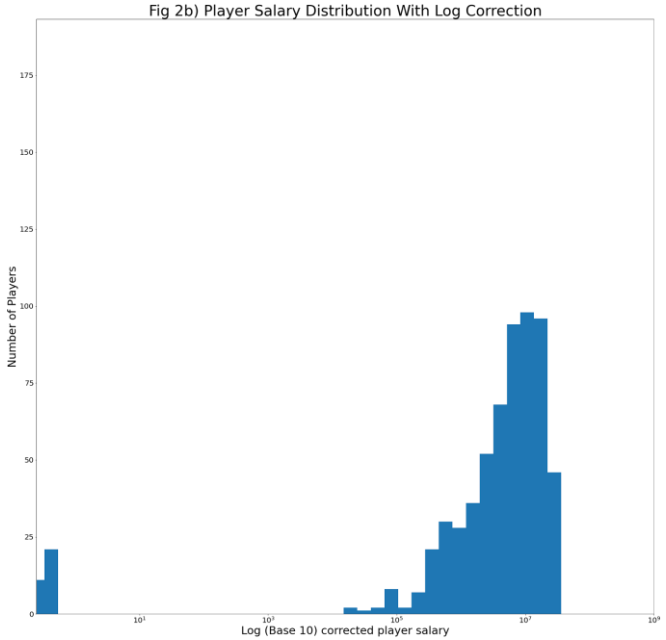
Fig 1) Correlation Heatmap



- Paneled grid = high correrlation between NBA stats in 1st, 2nd, 3rd yrs; can focus on 3x3 correlations
- Height and Weight positively correlated to each other and to rebounds; inversely correlated to avg field goal distance
- % field goals inversely correlated with FG distance
- Defensive Rating inversely corr to rebounds: future food for thought

Exploratory Data Analysis: Histograms (skew)





- Mean Annual Salary = incredibly right skewed
- Applied log transformation for gaussian distribution
- Generated two clear clusters:
  - 1st cluster = players not extended
  - 2nd cluster = Normally distributed with mean AAV of USD 10 mil
- Decided to do two models – 1) classify players into those who will remain in league vs. not; 2) Predict In-Prime AAV of players who remain.

Model 1: Classify Players Who Remain in League

- Fit decision tree to explore which traits most important
  - Right skewed data = predicted all players remain
- Upsampled data for model that attempted to learn valuable decision boundary
  - Resulting classifier much more balanced

Fig 4) Decision Tree without Upsampled Data

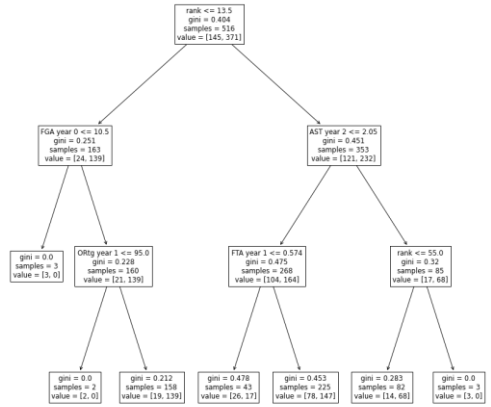
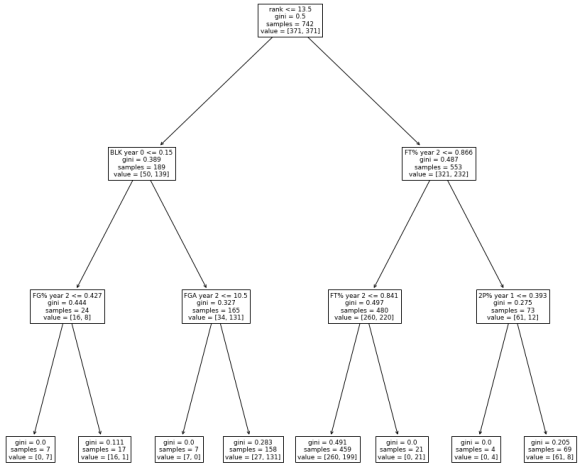


Fig 5) Decision Tree with Upsampled Data



Model 1: Classify Players Who Remain in League (contd.)

- Ran multiple models with various hyperparameters – logistic regression, bagging, boosting, support vector classifiers
  - Baseline accuracy per reference paper = 60%
  - Support vector machines performed worst
  - Best bagging and boosting models achieved 68% accuracy
- Traits that contributed most = draft rank, free throw %age, field goal %age, total rebounds

Fig 7) Bagging

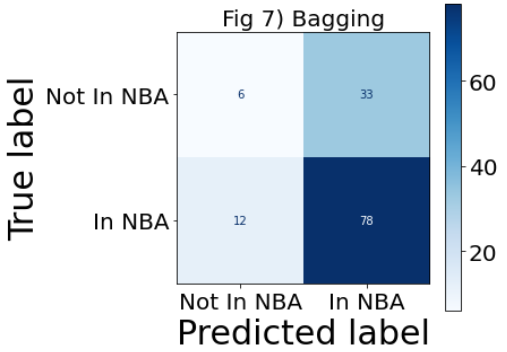


Fig 8) Boosting Model

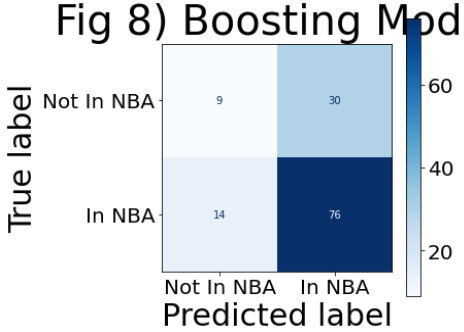
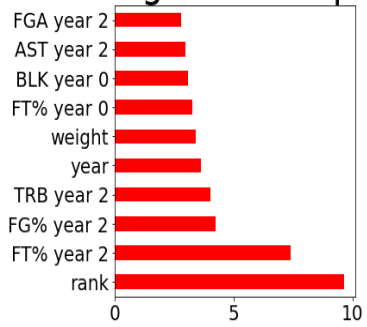


Fig 9) Boosting Model Important Traits



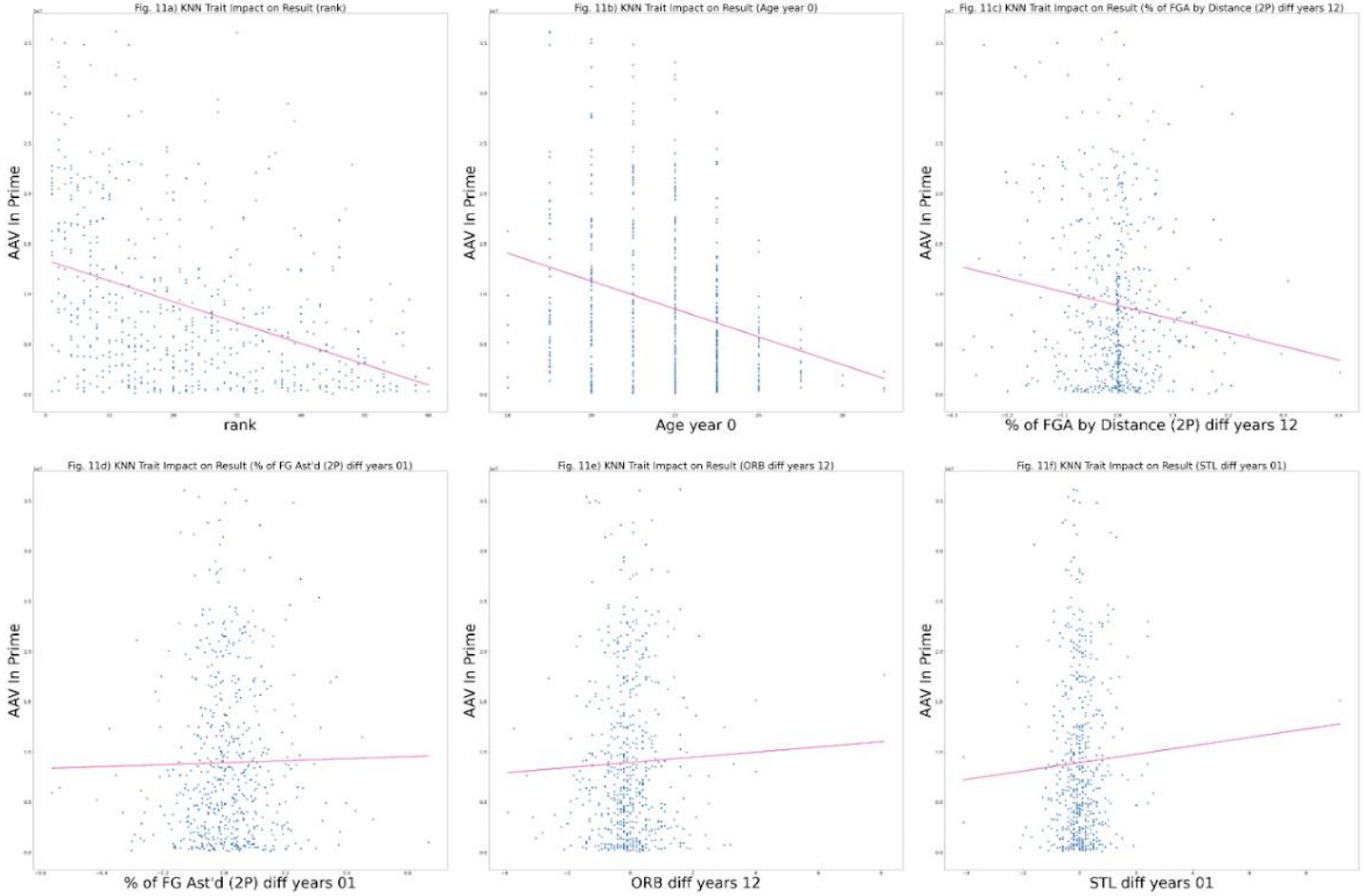
Model 2: Predict In-Prime AAV of players who remain

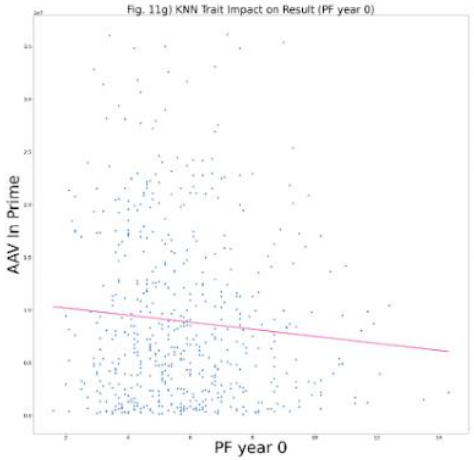
- Built linear regression with all traits to set floor
  - RMSE of $8.1 mil

- Performed lasso regression to eliminate traits that do not meaningfully contribute
  - Significant traits 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

Model 2 Findings (contd.)





- Draft rank, rookie age established before
- Interesting that younger age draft picks=higher value
- Interestingly, fouls inversely correlated to value

Future Extension

- Incorporate same data manipulation strategy of rookie year metrics together with growth rather than absolute values into classification model and assess accuracy.

- Scrape advanced statistics, such as RAPTOR data or value over replacement player, combine performance data, pre-NBA statistics to improve predictive performance of Models.