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

Krish Khanna

MPCS 53120 Applied Data Analysis

January 31, 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

Proposed Solution and Impact

- Predict Average Annual Salary in Prime (27-28 yrs) (=Dependent Variable)
- Multiple sets of indicators as independent variables to normalize factors outside the player's control
  - Combine Data (e.g., body fat %, age, sprint speed, bench press,…)
  - Performance Stats data:
    - College Performance
    - International Player Team Data, e.g., European leagues
    - NBA Statistics (Box score, Advanced Stats, etc.)
  - Impact:
    - Help original team decide whether rookie is valuable enough to extend
    - Serve as price ceiling for the team in negotiations so they do not overpay.

Data and Analysis

10 yr data for list of draft picks from Kaggle:

- Avg annual salary in prime: Web Scraper from ESPN data
- Combine data: scraped from NBA website
- College Performance: Web scraper for Sports Reference College Basketball
- International (pre-NBA): Web scraper for basketball-reference.com
- NBA Stats: Web scraper for NBA.com, and RAPTOR data set from fivethirtyeight.com

Proposed Analysis:

- Data Preprocessing:
  - Correlations for collinearity; Histograms to study distributions, remove outliers; whittle down useful statistics; processes for feature selection & hyper parameter tuning
- Iteratively fit regression models :
  Linear regression, decision tree regression, random forest regression, artificial neural networks, support vector regression, XGBoost regression, AdaBoost regression

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.

Takeaways

- Examine collinearity and think through how to deal with it
- Think about using voting model for final predictions
- Evaluate examine whether any model exhibits marked improvement over a pure minutes linear regression model
- Think about NaN/missing values; keep college, international, NBA data separate

Perceived Challenges and Action Plan

Plan

- Develop web scrapers for international, college performance, NBA stats, Annual Salary
- Python code to combine data
- Preprocessing: collinearity, outliers, distribution, etc
- Standard model creation, training, testing pipeline
- Iteratively fit regression models
- Evaluate model's performance: Mean salary prediction model, pure minutes played linear regression, 80-20 split

Challenges

- Need to look for new data sources if insufficient data post-scraping
- Collinearity and how to deal with it
- NaN/missing values; keep college, NBA, international data separate