Machine Learning in Basketball

By Aren Simmons

Predicting NBA Rookie Salaries

Nailing the NBA draft can result in value just about anywhere on the board. Most of the value however comes from evaluating what a player is truly worth.

The highest paid rookies are those taken first.

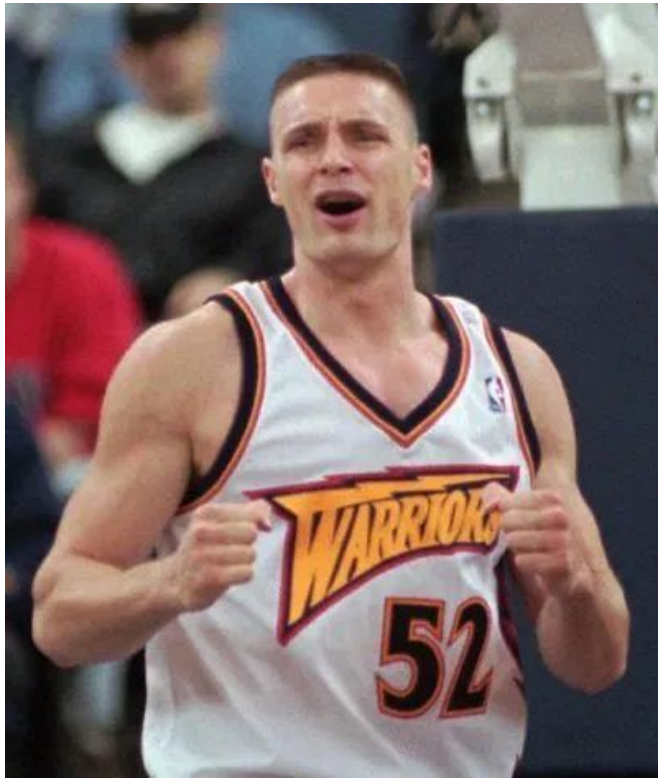
Many analysts reference college production, however does that necessarily translate to production in the NBA?

College stars who just didn’t pan out

Adam Morrison

- Picked 3rd overall by Charlotte Bobcats in 2006
- Made $3,616,680 ($4,565,733 after inflation) rookie year
- Played 4 seasons
- Averaged 7.5 pts, 2.1 rebounds, 1.4 assists

Todd Fuller



- Drafted No. 11 in 1996
- Made $1,125,000 ($1,838,933 adjusted for inflation) rookie year
- Never averaged more than 5 points
- Only lasted 5 years in the NBA

Jonny Flynn

- Picked 6th overall by the Timberwolves in 2009
- Made $2,969,280 ($3,526,124 adjusted for inflation) rookie year
- Played 3 years in the NBA
- Averaged 9 points, 3 Assists per game

Avoiding NBA Busts

Misallocation of a team's salary cap can significantly hinder that team's performance for years.

College statistics is only one of many factors that go into the team's decision to invest in a rookie player, however is it a good metric to use to evaluate how much to invest in players? If so, how accurate do teams nail these decisions?

To truly evaluate college production, we must dive into relationships between the stats and how good of predictors they are for salaries.

Data Acquisition

Two sources of data - NBA Salary (from data.world) and College Statistics (from Sports-reference.com)

NBA Salary was pulled as SQL query (12,000 rows)

College statistics required a web-scraper (5,000 rows)

Both were saved as csv files

Data Wrangling

Matched players across both datasets by player name and year.

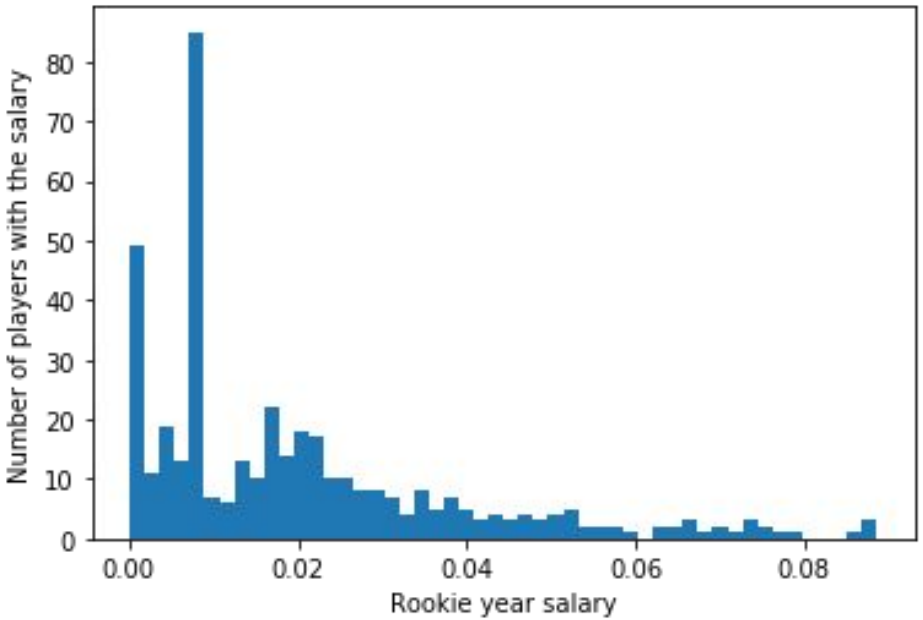
For NBA salaries, data was grouped by name and min. year to establish rookie year.

For College statistics, data was grouped by name and max. year to establish last year played in college before getting drafted.

Performed an inner join between the two datasets.

College stats were contextualized on per game or per min averages.

Salary Distribution



The distribution of salaries (expressed as % of salary cap) are skewed to the right.

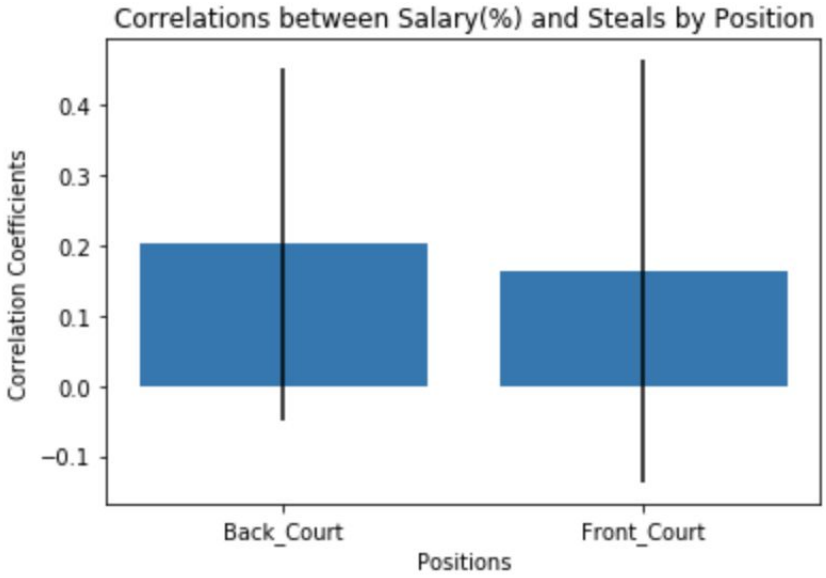
There must be notable stats that set these players apart.

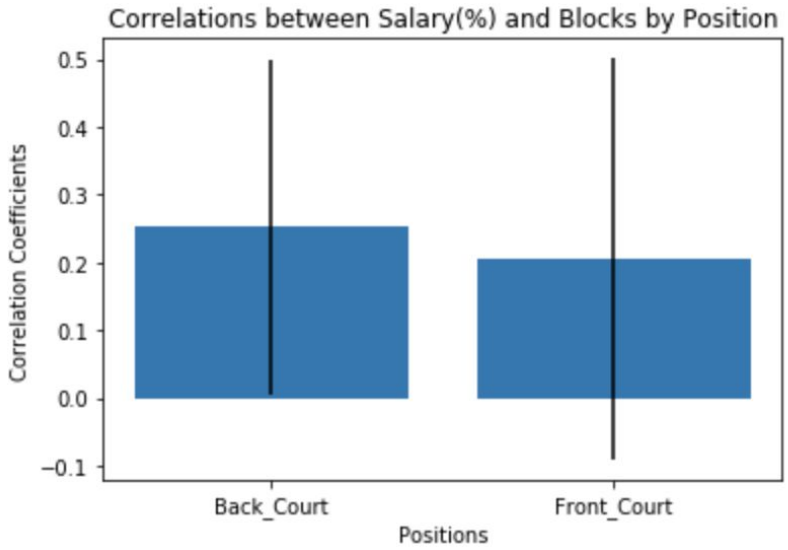
Exploratory Data Analysis

Distribution of Salaries by Class

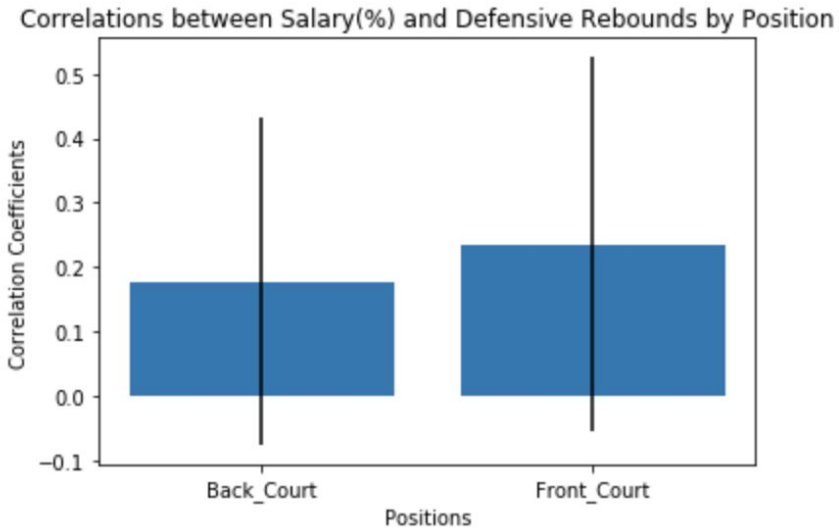
The variation of salaries is more evenly spread out for players who spend less time in college.

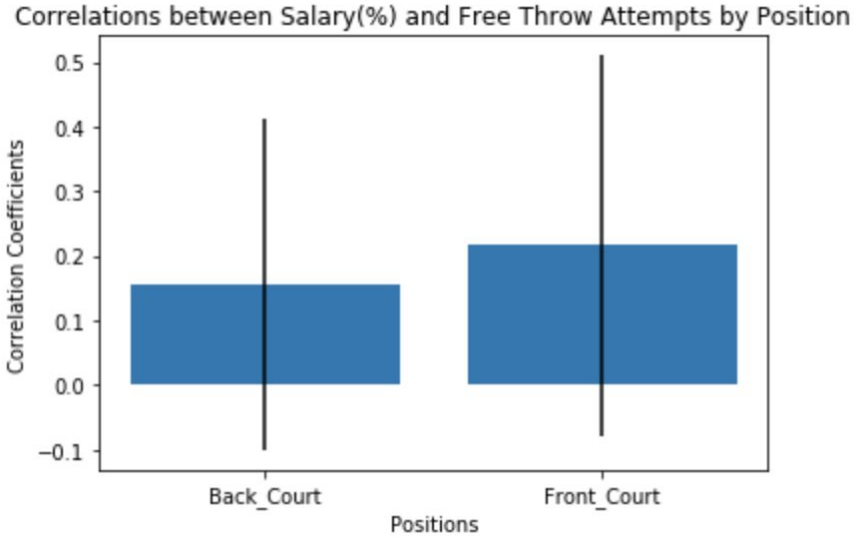
Blocks and Steals stat correlations





Defensive rebounds and FTA stat correlations





Positional Differences

A deeper dive reveals that depending on position there are stats that are higher correlated with salary:

I've divided these into front-court and back-court positions:

Back court (Guards) - Blocks and steals

Front court (Forwards & Centers) - Defensive Rebounds (DRB) and Free Throw Attempts (FTA)

The correlation between stats and salary separated by position are significantly different from one another (p-value <= .05).

Note: Because the number of observations are small ($n_{front\text{-}court} = 233$) & ($n_{back\text{-}court} = 188$) the margin of error is large.

Significant College Stats

After computing a correlation matrix, I've found that the stats that were most important across all positions in terms of being the highest correlated with salary (expressed as % of cap) were:

- Class (latest year played before draft)
- Field goals
- Field Goal Attempts
- 3 Pointers
- Free Throw Attempts
- Defensive Rebounds
- Steals
- Blocks

Model predictions

Different regression models were used to predict salary based on college stats:

Linear Regression

Ridge Regression

KNeighborsRegressor

Random Forest

Ensemble Ridge Regression 1 - Separated by position

Ensemble Ridge Regression 2 - Separated by position only for significant stats

Regression results (r-squared)

Linear Regression = 0.162

Ridge Regression = 0.169

Random Forest = 0.076

KNeighborsRegressor = .022

Ensemble Ridge Regression 1 = 0.167

Ensemble Ridge Regression 2 = 0.157

Conclusion

The models that predicted NBA rookie salaries were not very conclusive (~17% accurate).

College stats alone are not very good predictors of NBA rookie salaries.

There are a multitude of possibilities, however it is very hard to predict and comes down to the subjective nature of drafting.