Deal or No Deal?

An NBA Recommender System for Team Composition and Salary Optimization

Mathew Chandy, Leo Cheng, Lauren Okamoto
Mentored by Maksim Horowitz

TABLE OF CONTENTS

01

Intro & Data

Motivation

- Teams want to optimize their cap space, so it is helpful to know how much they “should” pay a player.
- Salary for the latest season is perhaps the best proxy for how much a player will demand for the upcoming season.
- Teams want a way to see if a player is over-valued based on current output-based determinants of salary
- Also need to account for a team’s unique needs

Goals

1. Estimate which players are over or under-valued by predicting salary surplus in dollars.

2. Evaluate how players fit with each other by predicting the probabilities of events that lead to an increase or decrease in expected points.

3. Provide a list of recommended players to add to a four-man lineup based on salary surplus and how much a hypothetical team is willing to pay a player, whilst accounting for complimentary playstyles.

Data Cleaning

Cleaning the NBA salary/stats dataset

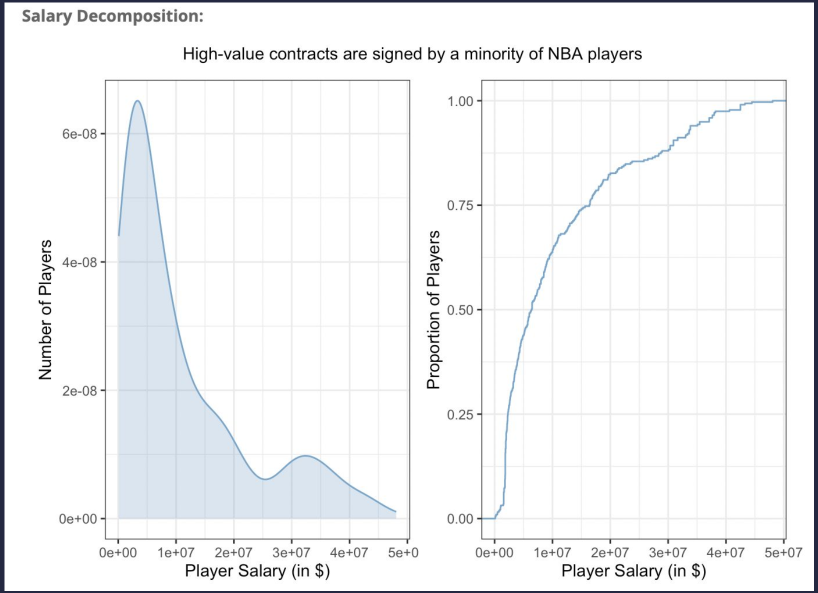
- Gathered salary and performance statistics data from the 2022 - 2023 season
- Removed redundant rows ensuing from Basketball-Reference's practice of listing all teams a player has been associated with
- Excluded players who averaged < 12 minutes per game
- Log-transformed player salary to normalize the skewed distribution and reduce the influence of outliers
- In the end: 317 observations

Cleaning the NBA play-by-play dataset

- Excluded players who were not in the top 250 players with the most possessions
- Got rid of "garbage time" possessions
- Limited free-throw situations to those where a shooting foul was drawn
- Used subsets of data to answer conditional probability questions
- In the end: 60,401 observations

Exploratory Data Analysis

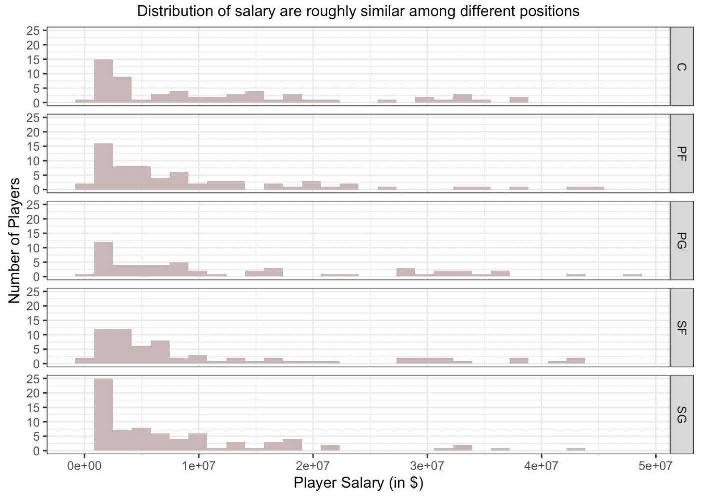
Are contract values equally distributed among players in NBA?



Exploratory Data Analysis

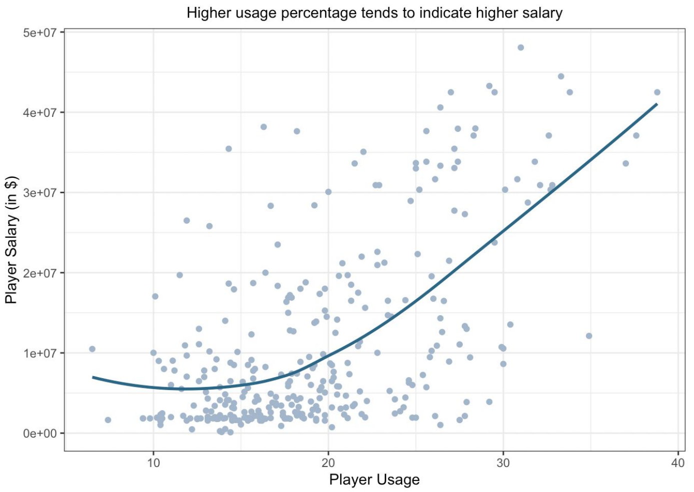
Do players in different positions get different payroll?

Salary Distribution by Position:



Do higher player usage lead to higher salary?

Salary Against Usage Percentage:



02

Methods

Random Forest Prediction Interval

Modeling Advantage:

Averaging variable outputs to improve prediction accuracy and control over-fitting

Prediction Output

- Generate centered interval predictions with 1/4 standard errors width
- Exponentiate both the lower and upper bounds to revert log transformation
- Subtract true salary of players' contract from the anticipated salary to compute for surplus values

Training Variable

- Games
- Minutes Per Game
- Years in League
- Two-point and Three-point Field-goals
- Free Throws
- Offensive and Defensive Rebounds
- Assists and Turnovers
- Steals and Blocks
- Points Per 100
- True Shooting Percentage
- Assist Percentage
- Usage Percentage
- Defensive Win Shares
- Offensive Box-Plus-Minus

Clustering

Purpose

Investigate whether there exists significant gaps in player contract and surplus values among different player archetypes

Gaussian Mixture Model

- Constructing player archetypes
- Observing & comparing performance-based statistics
- Categorizing NBA players into clusters

Modeling Result

- A VVE (ellipsoidal, equal orientation) model is selected with 3 components
- Cluster size : 64, 90, 163



How Can We Begin to Predict How a New Lineup May Perform?

Assessing Novel Lineup Performance via Similarity to Past Lineups

- For a general idea on how a novel lineup may perform, we minimize the Euclidean Pairwise Distance between our proposed lineup and a lineup from the 2022-2023 season based on 3 self-defined Player Metrics

- Limited 2022-2023 lineups to the 1000 most frequent ones

Metrics Used:

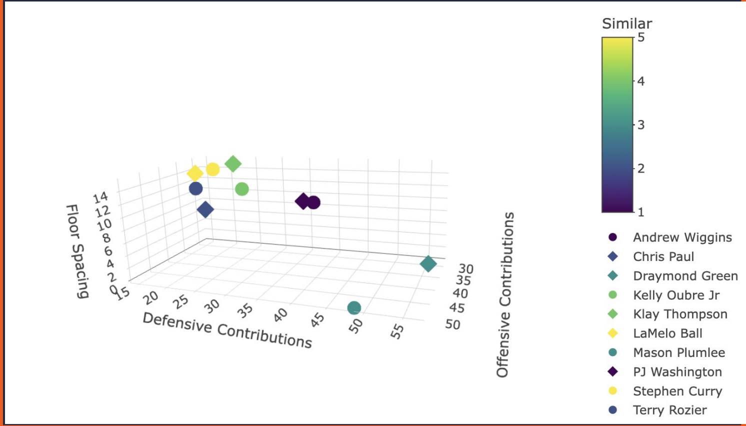
1. Offensive Contribution
2. Defensive Contribution
3. Floor Spacing

Proposed Lineup:

- Stephen Curry, Draymond Green, Andrew Wiggins, Klay Thompson, Chris Paul

Most Similar Lineup:

- LaMelo Ball, Mason Plumlee, PJ Washington, Kelly Oubre Jr, Terry Rozier



Multinomial Logit Model Event Tree

What Model?

- For our purposes we elected to use a multinomial logistic ridge regression model to predict the probability of the occurrence of each of four possession initiating events described in the Event Tree. We then used similar models to then predict the probability of a make, miss, foul, etc

- Regularization via Ridge Regression was necessary as our data has high multicollinearity due to the fact that many players play most of their minutes with the same subset of players

Fitting the Model/Predictors:

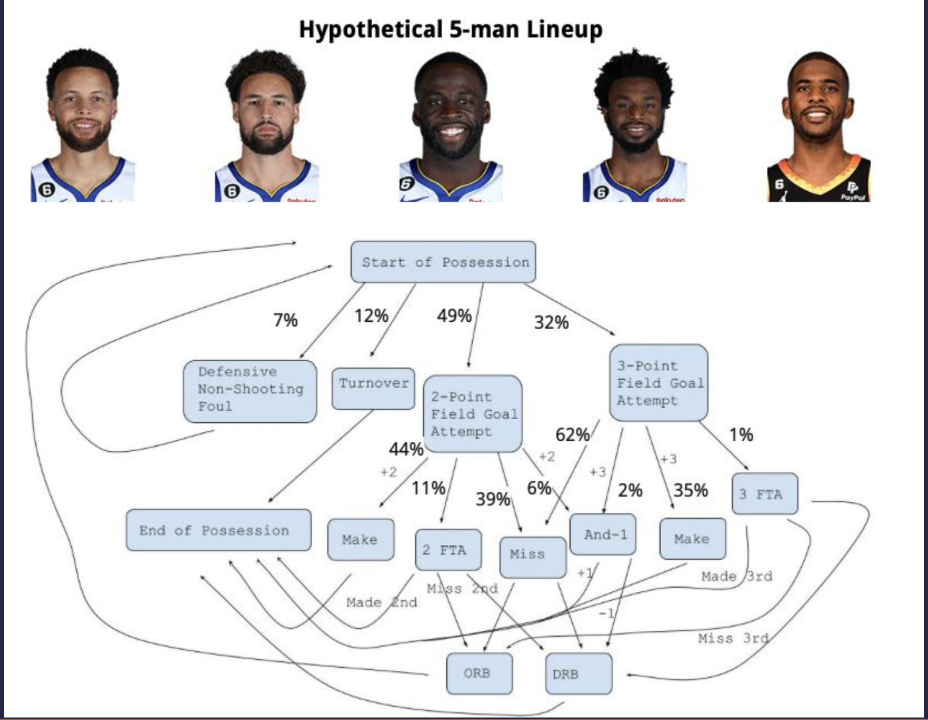
- Our Models were trained on data consisting of 500 indicators. 250 indicators represent if a player was on offense, and 250 indicators represent if a player was on defense.

- Our target variable is a the outcome for a particular situation

- When attempting to predict outcomes for a 5-man lineup on offense we set the respective player’s offensive indicators to 1, and every other indicator to 0

- When attempting to predict outcomes for a 5-man lineup on defense we set the respective player’s defensive indicators to 1, and every other indicator to 0

EVENT TREE


Hypothetical 5-man Lineup
Start of Possession
7%
12%
49%
32%
Defensive Non-Shooting Foul
Turnover
2-Point Field Goal Attempt
3-Point Field Goal Attempt
44%
+2
11%
39%
6%
+2
62%
+3
2%
+3
35%
1%
End of Possession
Make
2 FTA
Miss
And-1
Make
3 FTA
Made 2nd
Miss 2nd
+1
-1
Made 3rd
Miss 3rd
ORB
DRB

End Goal

Using the probabilities and the points generated from all these outcomes, we can predict an expected points when a five-man squad is on offense, an expected points when a five-man squad is on defense, and a net expected points by finding the difference of the two.

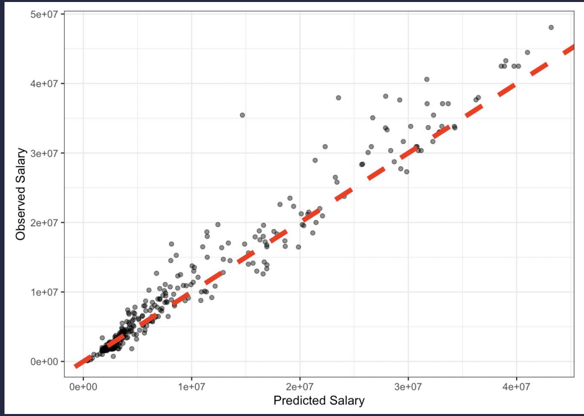
03

Results

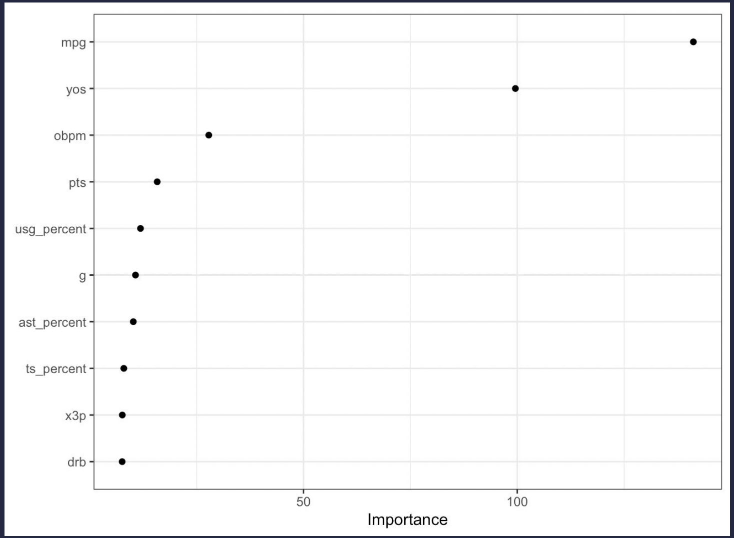
Random Forest Prediction Interval

Regression Result

Out-of-bag prediction error	●●●○○	0.4486
R-squared	●●●●○	0.6405



Variable Importance



Random Forest Prediction Interval

Assessment of Players' Cost-Effectiveness

Surplus = Predicted Salary - True Salary

Player List Preview

- *Ranked by contract surplus:*

```
##              player  g x3p x2p  ft orb  drb  ast stl blk tov  pts ts_percent
## 1  Kelly Oubre Jr. 48 3.3 7.5 4.9 2.0  5.7  1.7 2.1 0.6 2.0 29.9      0.534
## 2  Jordan Clarkson 61 3.7 7.3 4.8 1.7  4.2  6.5 0.8 0.3 4.5 30.5      0.558
## 3  Lauri Markkanen 66 4.2 7.8 7.3 2.7  9.2  2.6 0.9 0.8 2.7 35.5      0.640
## 4       Brook Lopez 78 2.7 6.9 3.0 3.2  7.3  2.0 0.7 3.9 2.2 24.9      0.630
## 5        Kyle Kuzma 64 3.5 7.7 3.8 1.2  8.9  5.2 0.8 0.6 4.1 29.5      0.544
## 6 Domantas Sabonis 79 0.5 9.5 5.7 4.4 12.6 10.0 1.1 0.7 4.0 26.4      0.668
```

- *Ranked by contract surplus lower bound:*

```
##              player  g x3p x2p  ft orb drb  ast stl blk tov  pts ts_percent
## 1  Dennis Schroder 66 1.8 4.7 5.2 0.5 3.4  7.1 1.2 0.2 2.7 19.8      0.545
## 2  Kelly Oubre Jr. 48 3.3 7.5 4.9 2.0 5.7  1.7 2.1 0.6 2.0 29.9      0.534
## 3        Kris Dunn 22 1.4 8.3 3.4 0.8 7.7 10.4 2.1 0.8 2.9 24.4      0.606
## 4  Jordan Clarkson 61 3.7 7.3 4.8 1.7 4.2  6.5 0.8 0.3 4.5 30.5      0.558
## 5    Royce O'Neale 76 3.3 1.3 1.0 1.1 6.7  5.7 1.3 1.0 2.3 13.6      0.538
## 6 Kevin Porter Jr. 59 3.4 6.0 5.0 1.8 5.7  8.1 2.0 0.4 4.5 27.1      0.565
```

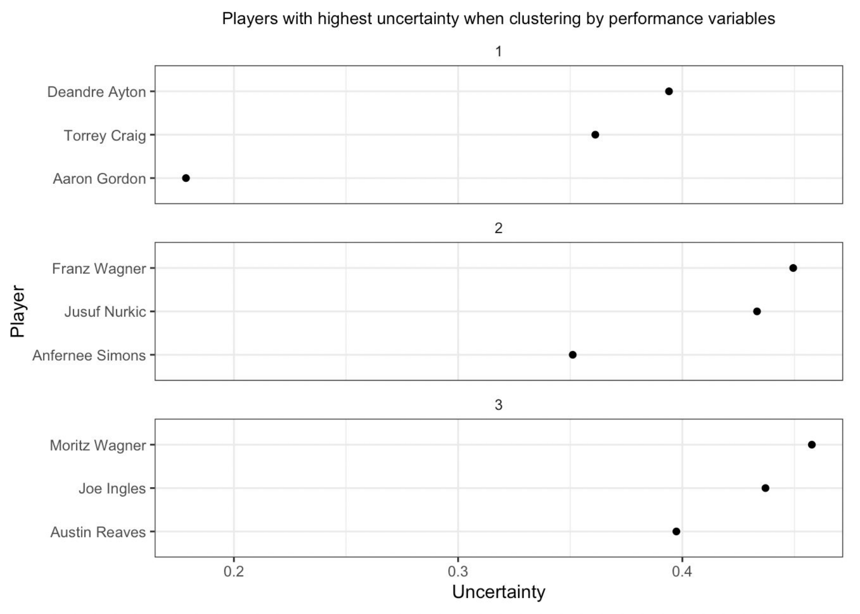
- *Ranked by contract surplus upper bound:*

```
##              player  g x3p  x2p  ft orb  drb  ast stl blk tov  pts ts_percent
## 1    DeMar DeRozan 74 0.8 11.1 8.3 0.6  5.6  6.8 1.5 0.7 2.8 33.0      0.592
## 2  Kelly Oubre Jr. 48 3.3  7.5 4.9 2.0  5.7  1.7 2.1 0.6 2.0 29.9      0.534
## 3  Lauri Markkanen 66 4.2  7.8 7.3 2.7  9.2  2.6 0.9 0.8 2.7 35.5      0.640
## 4  Jordan Clarkson 61 3.7  7.3 4.8 1.7  4.2  6.5 0.8 0.3 4.5 30.5      0.558
## 5 Domantas Sabonis 79 0.5  9.5 5.7 4.4 12.6 10.0 1.1 0.7 4.0 26.4      0.668
## 6    Jalen Brunson 68 2.8  9.4 6.8 0.8  4.2  8.7 1.3 0.3 3.0 33.9      0.597
```

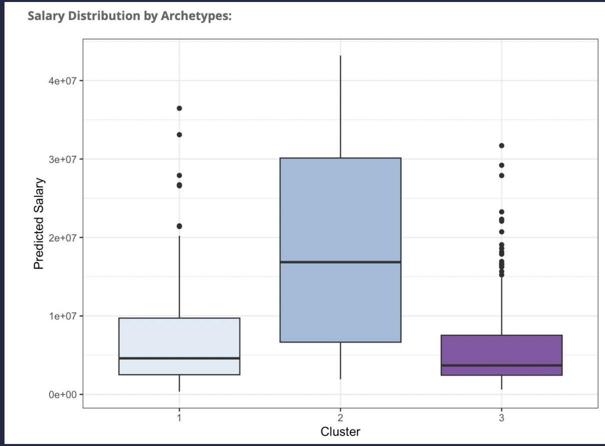
Clustering & Archetypes

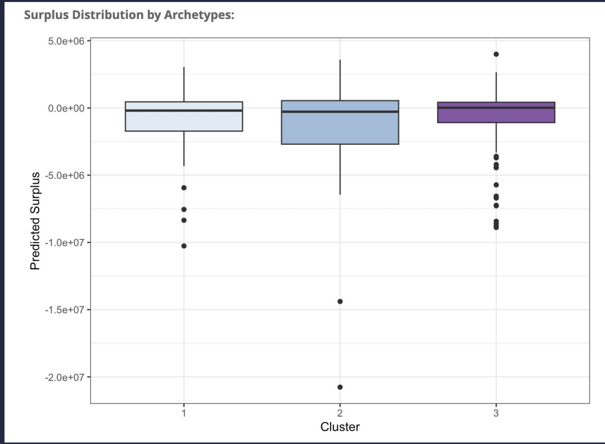
- **Traditional Bigs**
  - Highest total rebounds & blocks
  - Highest true shooting %
  - e.g. Myles Turner
- **Primary Scorers & Initiators**
  - Highest points, assists, usage, free throws
  - Highest offensive & defensive contribution
  - e.g. Lebron James
- **Roleplayers**
  - Efficient shooting
  - Highest game attendance
  - e.g. Tobias Harris

Model Uncertainty

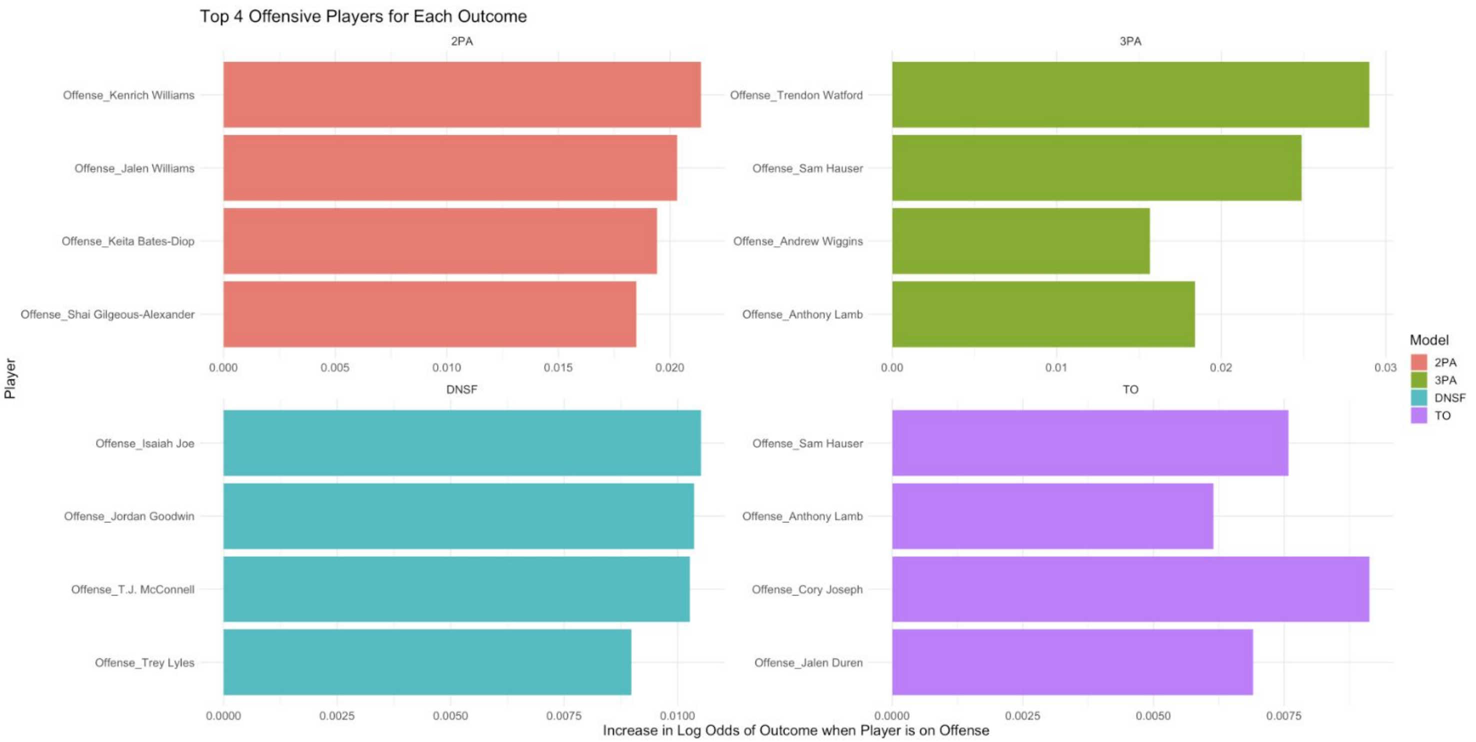


Clustering & Archetypes

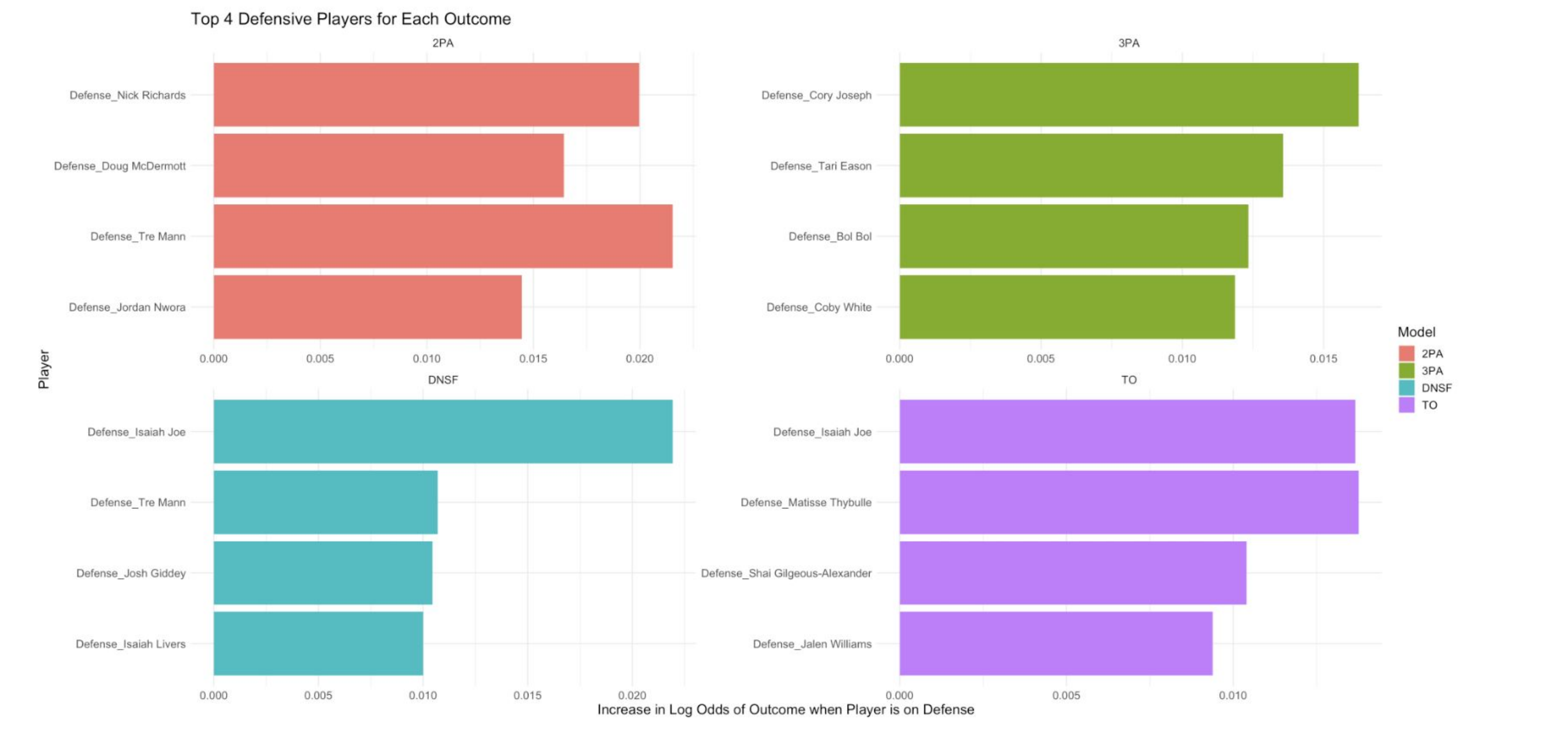




Multinomial Logit Model



Multinomial Logit Model

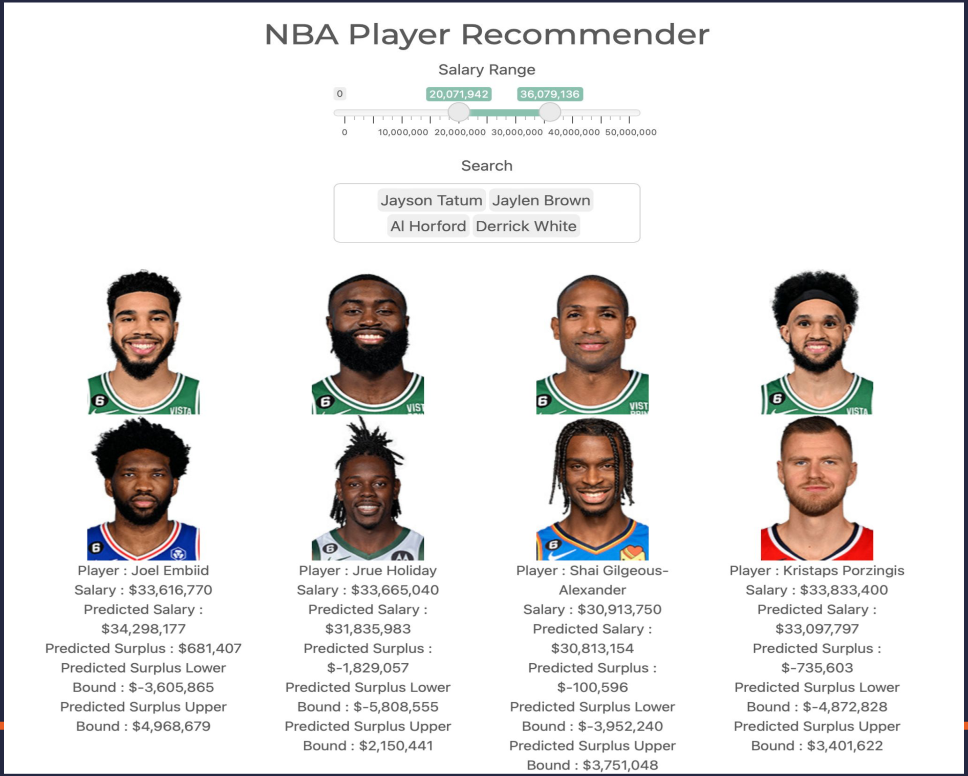


SHINY APP

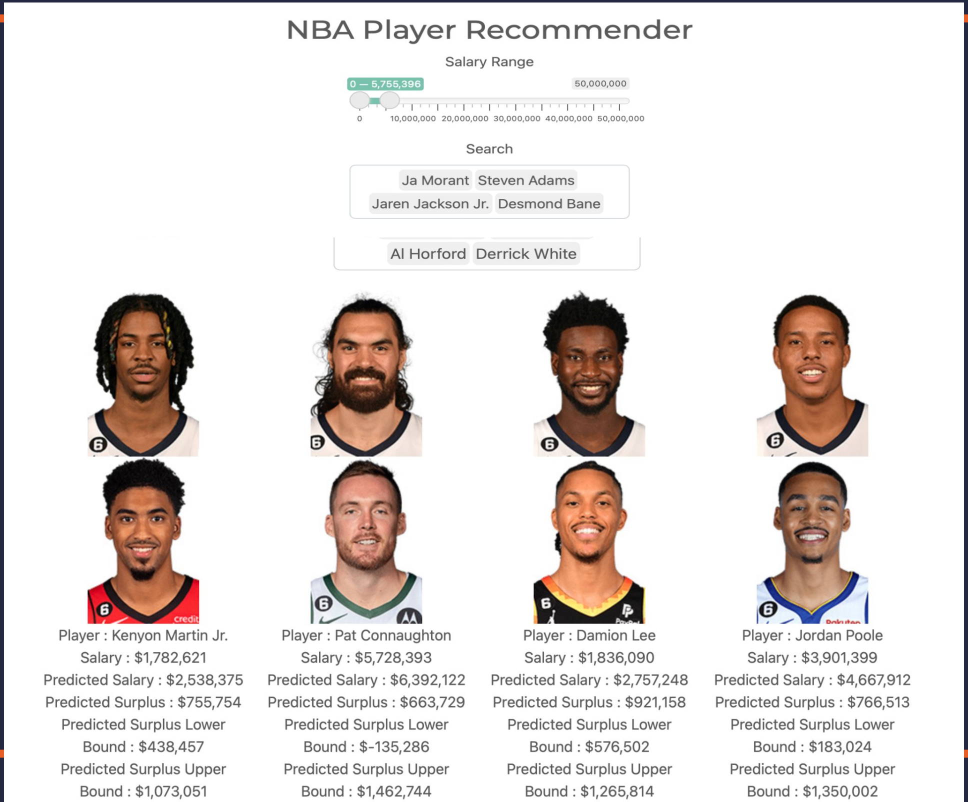
Our app takes five inputs: an interval of salary as a hard-cap and four players as members of a hypothetical 5-man lineup.

- First, the app filters out players whose 2022-2023 salaries fall out of the suggested salary interval.
- Then it sorts players based on the predicted surplus value of their contract in descending order and includes players whose indicated surplus is ranked in the top 20.
    - All three columns that account for players' surplus will be displayed : including a point estimate, a lower bound, and an upper bound
- Finally the expected points model is used to fit on the remaining player list and select the players with the highest expected points given that the 4 player inputs are members of the 5-man squad.

SHINY APP


NBA Player Recommender
Salary Range
0
20,071,942
36,079,136
0 10,000,000 20,000,000 30,000,000 40,000,000 50,000,000
Search
Jayson Tatum
Jaylen Brown
Al Horford
Derrick White
Player : Joel Embiid
Salary : $33,616,770
Predicted Salary : $34,298,177
Predicted Surplus : $681,407
Predicted Surplus Lower Bound : $-3,605,865
Predicted Surplus Upper Bound : $4,968,679
Player : Jrue Holiday
Salary : $33,665,040
Predicted Salary : $31,835,983
Predicted Surplus : $-1,829,057
Predicted Surplus Lower Bound : $-5,808,555
Predicted Surplus Upper Bound : $2,150,441
Player : Shai Gilgeous-Alexander
Salary : $30,913,750
Predicted Salary : $30,813,154
Predicted Surplus : $-100,596
Predicted Surplus Lower Bound : $-3,952,240
Predicted Surplus Upper Bound : $3,751,048
Player : Kristaps Porzingis
Salary : $33,833,400
Predicted Salary : $33,097,797
Predicted Surplus : $-735,603
Predicted Surplus Lower Bound : $-4,872,828
Predicted Surplus Upper Bound : $3,401,622

SHINY APP


NBA Player Recommender
Salary Range
0 — 5,755,396
50,000,000
0
10,000,000 20,000,000 30,000,000 40,000,000 50,000,000
Search
Ja Morant Steven Adams
Jaren Jackson Jr. Desmond Bane
Al Horford Derrick White
Player : Kenyon Martin Jr.
Salary : $1,782,621
Predicted Salary : $2,538,375
Predicted Surplus : $755,754
Predicted Surplus Lower Bound : $438,457
Predicted Surplus Upper Bound : $1,073,051
Player : Pat Connaughton
Salary : $5,728,393
Predicted Salary : $6,392,122
Predicted Surplus : $663,729
Predicted Surplus Lower Bound : $-135,286
Predicted Surplus Upper Bound : $1,462,744
Player : Damion Lee
Salary : $1,836,090
Predicted Salary : $2,757,248
Predicted Surplus : $921,158
Predicted Surplus Lower Bound : $576,502
Predicted Surplus Upper Bound : $1,265,814
Player : Jordan Poole
Salary : $3,901,399
Predicted Salary : $4,667,912
Predicted Surplus : $766,513
Predicted Surplus Lower Bound : $183,024
Predicted Surplus Upper Bound : $1,350,002

04

Discussion

Discussion

Importance

1) Evaluate if a player is currently over or undervalued
2) See if a player is a good fit for the team.
3) The visualizations on salary & surplus distribution given constructed archetypes are noteworthy as they reveal a significant divergence in salary rates but barely any difference in surplus values among different player archetypes.

Limitations

- In terms of the clustering process, we didn't establish a clear dividing line between archetypes, and it relies on comparing average values and integrating knowledge of basketball.
- Small Sample Size and Collinearity
- Our model does not take into account which player got fouled, turned over the ball, or shot the ball.

Next Steps

- Involve more complexity and variability in modeling by training on data from various seasons
- Employ a better refined clustering model to construct more precise player archetypes
- Attempt to implement method which predicts outcomes on propensities of individual players to commit an action.

Questions?