Deal or No Deal ?
An NBA Recommender System for Team Composition and Salary Optimization

Mathew Chandy, Leo Cheng, Lauren Okamoto
Project Advisor: Ron Yurko, Maksim Horowitz

Background & Introduction

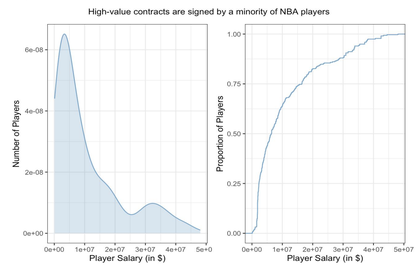
In the dynamic ecosystem of professional basketball, the ability to discern the true value of a player often lies at the intersection of performance on the court and the economic dynamics of the sport. By venturing beyond traditional evaluation metrics, we seek to shed light on the nuances of the game and offer insights that could redefine team strategies. Our investigation delves into whether player compensation aligns with their performance, revealing potential hidden gems. Recognizing that basketball is a team sport, we also aim to identify optimal player combinations for a hypothetical team. This challenge has been tackled previously by methods developed by Maymin [1] and Kuehn [2], but we seek to combine the two team goals of positive surplus value and complimentary playstyles into a single Shiny App. Through this project, we hope to offer a demonstration of a salary surplus model for an individual and an expected points model for a five-man lineup, the two of which can be expanded and strengthened with more data.

Data

Our project can be understood to be tackling two problems that each demanded different data.

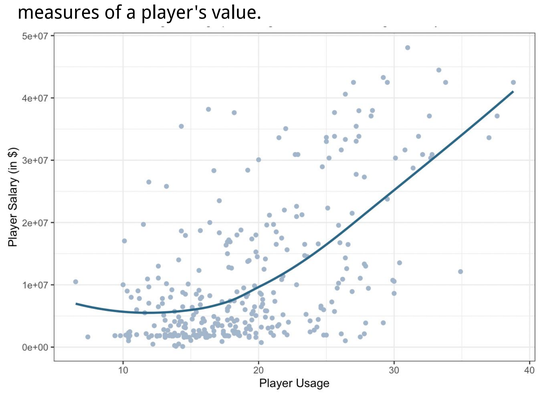
The first problem is a **salary surplus prediction model**.
A player's salary surplus can be expressed as:
Surplus = Predicted Salary - True Salary
Therefore, we need to predict salary and use the residuals to evaluate a player's surplus value.
Response Variable: Salary

Salary Decomposition:



Luckily, salaries for all players for the 2022-2023 season are publicly available. Once we extracted this information from Basketball Reference [3], we merged it with individual basic and advanced statistics data (also from Basketball Reference [4]), and excluded players who averaged less than 12 minutes per game.
In the end we were left with 317 observations.
Below are predictor categories of interest:
- Player information (Age, Years in League)
- How much a player plays
- Shooting per 100 possessions
- Other basic stats
- Advanced stats

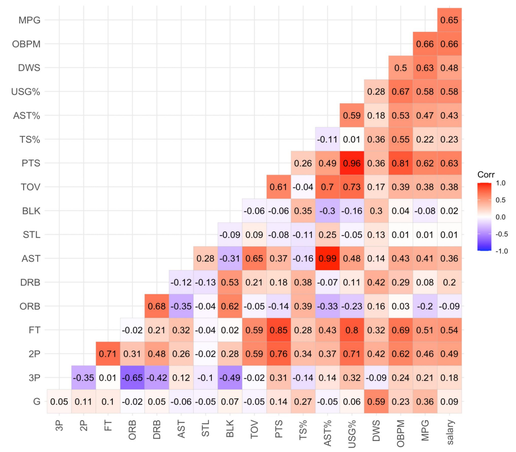
This data can be used to create a model predicting salary based off conventional performance-based measures of a player's value.



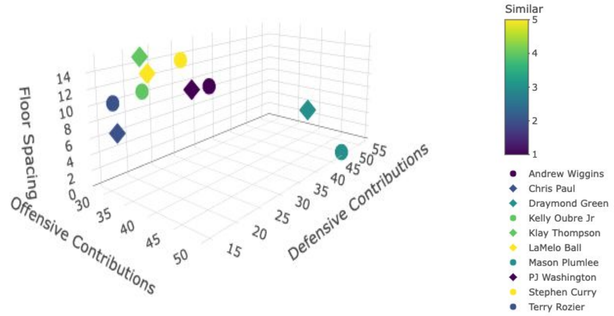
Higher Usage% → Higher Salary

The second problem is actually attempting to account for a team's need. To make it a little more approachable, we considered the simple question of if you had a team of 4 players, and you had to choose one more player to complete the squad, who would it be? To answer this question, we need play-by-play data to retrieve what was the outcome of a possession with a certain offensive and defensive lineup. This data for the 22-23 season was acquired from Ramiro Bentes' GitHub [5]. We included the 250 players with the most possessions and ignored garbage time possessions. In the end we were left with 60,401 observations.
For possession-level data we need:
- What the initial event for the possession was
- If a shot was taken, did they make it and were they fouled
- If they were fouled, how many shots did they make
- If a shot was missed, who got the rebound
- Indicator variables for who was on the court

This data can be used to create an event tree of models to measure **which players complement each others' playstyles**.



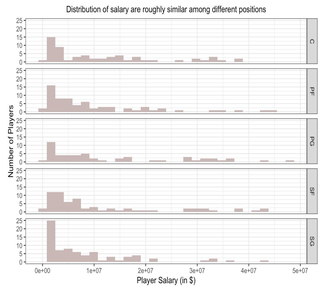
Using Scatter Plots to Assess Similarities Between Novel and Past Lineups
Proposed lineup of Andrew Wiggins, Stephen Curry, Draymond Green, Klay Thompson, and Chris Paul is most similar to a former Hornet's Lineup of Kelly Oubre Jr, PJ Washington, Lamelo Ball, Terry Rozier, and Mason Plumlee based euclidean distance given various player metrics. This tool serves as a general basis for beginning to understand how a brand new lineup may perform given their lineup comparison.



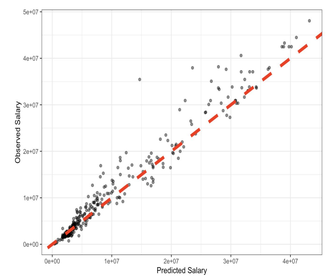
Analysis & Modeling

Random Forest Prediction Interval
To generate predictive salary results and compute surplus value of individual players, we used the Random Forest Algorithm for regression modeling, whose advantage lies in averaging variable outputs to improve prediction accuracy and control over-fitting. After eliminating irrelevant explanatory variables and shrinking the number of variables involved to minimize the influence of multicollinearity, the random forest model was trained based on predominantly performance-based statistics.

Salary Distribution by Position:

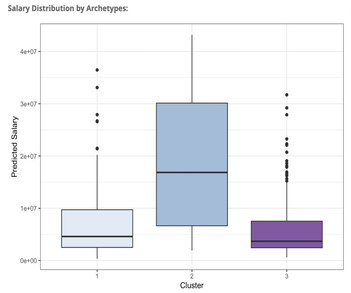


Model Fit:

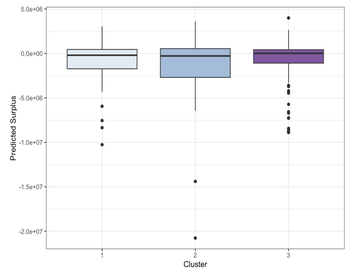


To provide a range of plausible values and reflect potential uncertainty in salary prediction, we modified the model to produce centered interval estimates with 1/4 standard errors width. In terms of interpreting the predictive results, we reverted the effects of log transformation by exponentiating the standard errors of our predictions. True salary of individual players was then subtracted from the predicted salary to compute for surplus values in potential contracts.

Salary Distribution by Archetypes:



Surplus Distribution by Archetypes:

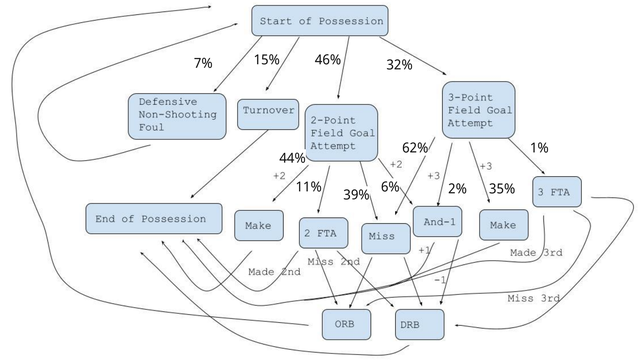


Multinomial Logit Model Tree for Expected Points per Possession
A model predicting surplus value does not take into account how well a certain player may fit with four other players. To address this, we sought to create a model that predicted the net expected points and given a lineup. To measure interaction between players, we need a model that does not merely predict the probability of good or bad outcomes. Given a lineup, our model predicts the probability of an event such as a 2-Point Field Goal Attempt, and then predicts the probability that the event will actually lead to a positive outcome: the shot being made, for example. This relies on 3 Multinomial Ridge Logit Models, and 1 Ridge Logit Model. This is similar to Maymin's [1] method, which instead used a Probit Model. Below is a table indicating the best MSE for the four Ridge Regressions.

Model P(EVENT|..)	λ Minimum MSE	Mean Squared Error
P(E|Start of Possession)	0.01132061	2.06629
P(E|2PFGA)	1.327846	2.297455
P(E|3PFGA)	10.90901	1.558885
P(E|MISS)	0.0436495	0.1641992

These probabilities can be used to estimate expected points when a hypothetical lineup is on offense and on defense.

Hypothetical 5-man Lineup



Conclusion & Takeaways

The methods in this study could prove to be very useful for NBA teams to inform trade decisions. This project provides a potential way for teams to 1) evaluate if a player is currently over or undervalued and 2) see if a player is a good fit for the team.
Discussion: Based on the traditional player position metric, we cannot easily conclude whether players are unproportionally paid depending on their play-styles. Therefore, 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: This project was limited by its scope, all of our models were only trained on data from the 2022-2023 NBA Season. Because of the small sample size, although we accounted for collinearity somewhat with ridge regression, it is still possible that our model will misassign contributions if a player was always on the court with other players. Our model also does not take into account which player initiated an action, rather measuring a player's contribution by mere presence on the court. In terms of the clustering process, we also did not establish a clear dividing line between archetypes, and it generally relies on comparing average values and integrating knowledge of basketball.
Next Steps: Train on more data, employ a more refined clustering model, and try implementing Kuehn's [2] method, which predicts outcomes on propensities of individual players to commit an action.

References:
[1] Maymin, A., Maymin, P., & Shen, E. (2013). NBA chemistry: Positive and negative synergies in basketball. International Journal of Computer Science in Sport, December. [2] Kuehn, J. (2017). Accounting for complementary skill sets: evaluating individual marginal value to a team in the National Basketball Association. Economic Inquiry, 55(3), 1556-1578. [3] <https://www.basketball-reference.com/contracts/players.html> [4] <https://www.basketball-reference.com/leagues/NBA_2023_per_poss.html>
[5] <https://github.com/ramirobentes/NBA-in-R/blob/master/2022_23/regseason/pbp/pbp_lineups.R>