Modeling & Predicting Optimal NBA Lineups

Sid Iyer, Sparsh Jain, Wesley Graham, Roshan Srinivasan, Nabeil Kizilbash
Machine Learning & Data Analytics - Fall 2019

Introduction & Motivation

In the world of the National Basketball Association (NBA) today, team executives spend far too much money on their teams. Over 20 years, the average amount spent on building team rosters has risen from $30M to a whopping $126.7M (this includes luxury tax on top of the NBA salary caps). Our team believes that there may be more informative methods and metrics that NBA team owners can analyze to be able to optimally choose their teams. Here, we investigate the world of fantasy leagues.

Data

We are essentially looking for the best combination of variables that help predict player efficiency. In this case, player efficiency is defined as fantasy points/salary. Based on this metric, we explore:

"How might we assemble an NBA team using the roster restrictions in order to optimize team performance and fantasy league points while still being under the salary cap?"

We picked our data from three sources: NBA (Fantasy Point Formula), ESPN (Salary), and Basketball Reference (Player Performance by Season). By building a simple scraper, we were able to pull the player name, team and corresponding salary for each year, and used the NBA fantasy league website to compute the fantasy point scores that corresponded to each player for the 2016-2019 seasons. Since the NBA's Fantasy League site is official, we believe that the data is most credible from here as compared to a third-party fantasy league website.

We then performed data merge based on primary key unique identifiers of players, in order to remove duplicates and matching players with different spellings and other inconsistencies. After the data merge & standardization, we then moved forward with analysis.

Analytics

Key Objective: To determine optimal player value based on total fantasy points & salary per year.

Response Variable = (Salary) / (Fantasy Points)

Here are the steps that were taken in order to begin modeling:

1. Split the data into a training set (2017-2019, 72.9%), & a test set (2016-2017, 27.1%). The 2018-2019 set has a variety of starting positions for some players (i.e. if a player starts both as a point guard and center over most of the games, their position is labeled as PG-C, so that they are not classified into any one of those groups).
2. Generated a response variable by dividing salary by fantasy points.
3. Removed all unnecessary columns like player name, ID, salary, and fantasy points.

We assumed our baseline model to assume that there is no predictive power in our model and that the optimal team selection is merely based on the fact that the lowest salary/fantasy points are probably the best players to choose. Hence, the OSR^2 for the baseline model is 0.

1. Attempted a random forest model using all the features (e.g. age, team, % of 3 pointers scored, free throw count etc.) and varied the mtry value. By using TuneRF in R, we tuned the mtry value to 3, nodesize to 5 and ntree to 300. The initial OSR^2 value was ~0.4, which showed that there is much greater room for improvement in the modeling.
2. As observed in Appendix A, we then used linear regression and VIF scores to perform optimal variable selection and removed any features that had extremely high correlation. It is important to note that X2P. factor which is the 2 point percentage of a player as a correlation value confidence interval that is less than 0, which would imply that it should be removed. However, because the factor's VIF score was less than 5, we chose to keep it in the feature set.
3. We then re-ran the random forests (using TuneRF) and cross-validation by varying mtry (15), ntree (1000) and nodesize (9) to achieve an OSR^2 score of ~0.6, a significant improvement from our initial model.
4. We also tried boosting using the entire feature set with the k-fold value = 5 for cross-validation. This gave us an OSR^2 = ~0.4 which was not that good.
5. In order to validate both models, we performed bootstrapping to understand the variability of the OSR^2 value obtains. Using a resample value = 10000, we obtained an interval of [0.4, 0.8] for Random Forest and [0.1, 0.6] for Boosting. This means that Boosting is a really bad model since it has high variability and it has a lower OSR^2. It is also important to note that there are other models that we could have tried such as blending or stacking. However, those models would have required validation sets, and unfortunately, our datasets were not big enough to generate them given that we were only able to obtain accurate player data for only 3 seasons. Additionally, for a model like blending, if we were to combine boosting with random forest, it would be fair to hypothesize that the OSR^2 would be lower than the optimum since we would be combining a model with a really high OSR^2 with a model with a lower value.
6. Lastly, we appended the predictions to the test set.

After running the models, we used the predictions in order to develop our optimal roster and see how our model's performance matches up with the next best alternative.

Roster Selection Criteria:

1. Maximum Roster Size = 15 Players
2. Minimum 2 Point Guards, 2 Shooting Guards, 2 Small Forwards, 2 Power Forwards and 2 Centers.
3. Maximum Working Salary = $109.14M for the whole team

In order to create our roster, we iterated through our table in descending order from most predicted points to least and asked the following questions: "is this player's salary lower than what salary we have remaining?" and "have we got enough of the required position?" After filling out all the required positions, we iterated through the list once again and picked whoever fit our remaining salary cap. We then terminated the loop once we filled our roster size.

Final Lineup:

Our final lineup (Appendix B) consists of a roughly equal mix of present day NBA starters and role players. What is interesting to note is that our roster currently consists of zero NBA all-stars - which is a discrepancy against our initial expectations. This potential lack of correlation between projected fantasy output and "fan popularity" could however be treated as an advantage to executives, as our team can be contracted for under $75M USD - (25% less than the salary cap for NBA executives).

In terms of team performance, our generated team on average produces 400% more predicted fantasy points than the average NBA player - and costs executives only $156 USD/fantasy point (vs $8,750 USD/fantasy point regularly). This on-court efficiency is better than the efficiency of any other present NBA team (Appendix C).

Notable Statistics

Metric	**League Average**	**Our Team**
Salary Cap	$109.14M	**$73.78M**
Points per Player	10,800	**40,512**
$/Point Spent	$8,750	**$156**

All in all, our optimal team roster generated a cumulative 607,680 points, whereas the next best NBA team's fantasy points were merely 146,445.

Impact

Increased usage of data in sports has resulted in a new era of more efficient decision-making and increased monetary gains. As an NBA owner, one would want to use as much data as possible, including crowdsourced metrics such as fantasy data, to ensure the highest ROI and ultimately success. However, given that this is a team sport, one caveat of our model is if it ultimately contributes to wins for a team. For example, would having the best players on a team cannibalize each player's stats and performance? How will that affect each player's value?

Aside from basketball, the ideas presented in this project can be applied to other sports and avenues of life. For example, a similar approach has been taken by several teams in Major League Baseball, most notably the Oakland Athletics, as seen in their "Moneyball" approach to finding effective players for less money. As a result, the A's have consistently fielded competitive teams every year, making the playoffs several times in the last decade.

The same ideas can also be applied to non-sport situations as well, including medicine (prescribing medicine more accurately), insurance (determining what the best insurance package to purchase would be for a family given their income levels), or even the political sphere (effective use of campaign funds to achieve success & efficiency).

In a world where data analysis is increasingly important, those who effectively can use data to their advantage to predict efficient outcomes will be the most successful.

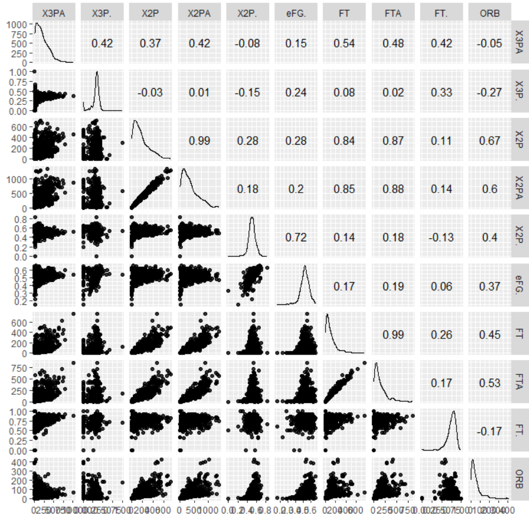
References

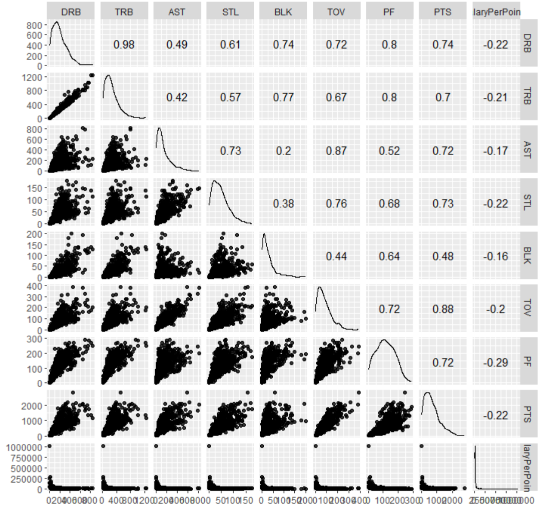
"Basketball Statistics and History." *Basketball*, www.basketball-reference.com/.

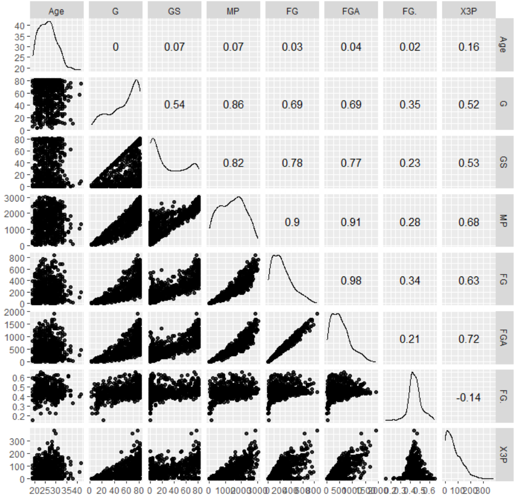
"NBA - National Basketball Association Teams, Scores, Stats, News, Standings, Rumors." *ESPN*, ESPN Internet Ventures, www.espn.com/nba/.

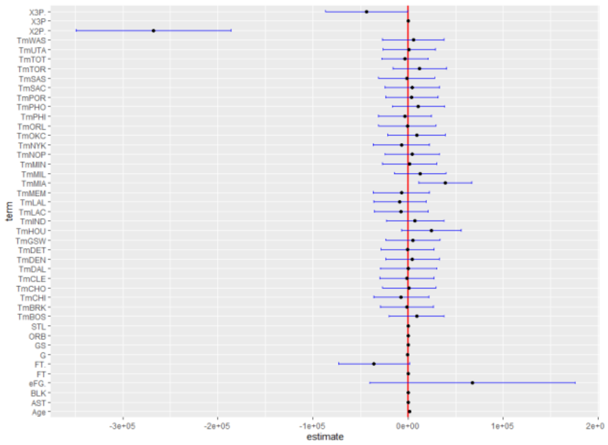
NBA.com. "The Official Site of the NBA." *NBA.com*, NBA.com, 24 Sept. 2019, www.nba.com/.

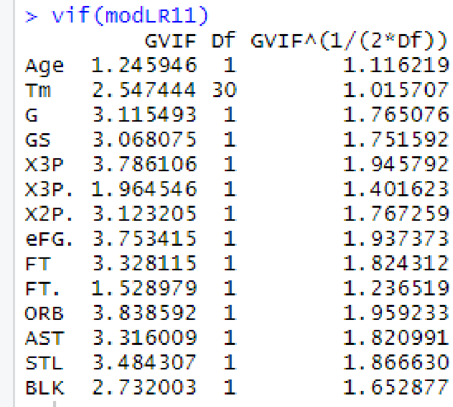
Appendix A





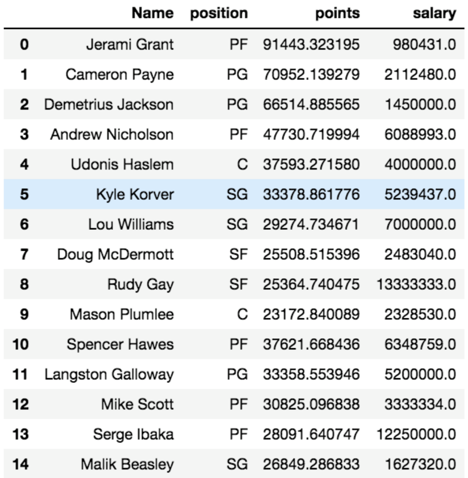






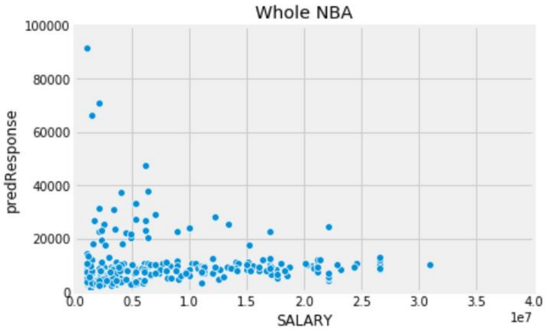
```
> vif(modLR11)
          GVIF Df GVIF^(1/(2*Df))
Age   1.245946  1        1.116219
Tm    2.547444 30        1.015707
G     3.115493  1        1.765076
GS    3.068075  1        1.751592
X3P   3.786106  1        1.945792
X3P.  1.964546  1        1.401623
X2P.  3.123205  1        1.767259
eFG.  3.753415  1        1.937373
FT    3.328115  1        1.824312
FT.   1.528979  1        1.236519
ORB   3.838592  1        1.959233
AST   3.316009  1        1.820991
STL   3.484307  1        1.866630
BLK   2.732003  1        1.652877
```

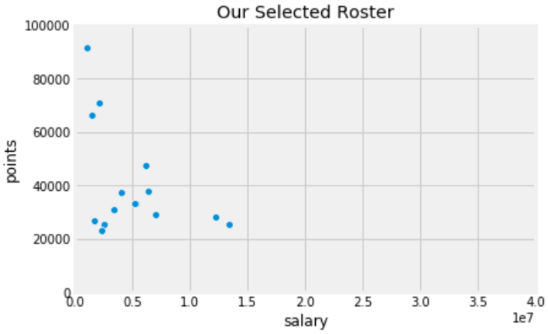
Appendix B

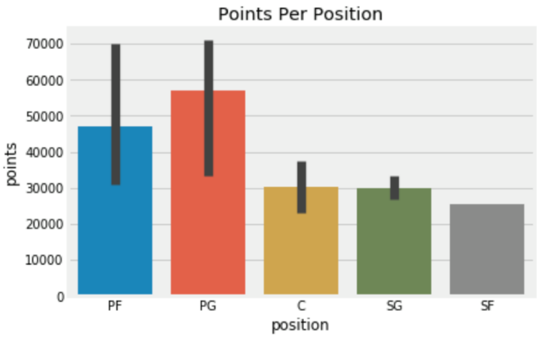


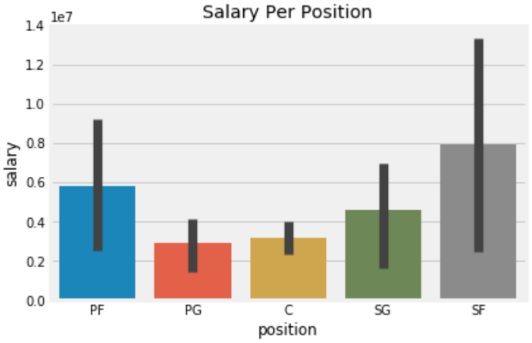
	Name	position	points	salary
0	Jerami Grant	PF	91443.323195	980431.0
1	Cameron Payne	PG	70952.139279	2112480.0
2	Demetrius Jackson	PG	66514.885565	1450000.0
3	Andrew Nicholson	PF	47730.719994	6088993.0
4	Udonis Haslem	C	37593.271580	4000000.0
5	Kyle Korver	SG	33378.861776	5239437.0
6	Lou Williams	SG	29274.734671	7000000.0
7	Doug McDermott	SF	25508.515396	2483040.0
8	Rudy Gay	SF	25364.740475	13333333.0
9	Mason Plumlee	C	23172.840089	2328530.0
10	Spencer Hawes	PF	37621.668436	6348759.0
11	Langston Galloway	PG	33358.553946	5200000.0
12	Mike Scott	PF	30825.096838	3333334.0
13	Serge Ibaka	PF	28091.640747	12250000.0
14	Malik Beasley	SG	26849.286833	1627320.0

Appendix C


Whole NBA
predResponse
1e7
SALARY


Our Selected Roster
points
1e7
salary


Points Per Position
points
PF
PG
C
SG
SF
position


Salary Per Position
1e7
salary
PF
PG
C
SG
SF
position