Finding the Top Undervalued NBA Players from 2017-2018

Jacob Meeker
Bellevue University, jmeeker@my365.bellevue.edu
20 November 2020

Business Problem

In any professional sports league, front office money management is a very important factor to the success of a team. Teams that utilize their salary caps the most efficiently are more likely to have a successful season. Because of this, my idea for this project was to find the most undervalued NBA players (from the 2017-2018 season). This could be significant for organizations to determine whether they can acquire a player at a good value or how much is worth spending on a player. This could also be used by players to negotiate for what they are worth. This analysis is relevant because it could help improve a current team as well as solidify the future of the team's success. As shown in Figure 1, over half of the NBA players from the 2017-2018 made less than $5 million per year. Being able to find good value in cheaper players could help push a title contender over the top, especially if they are near their max cap space (which contenders usually are).

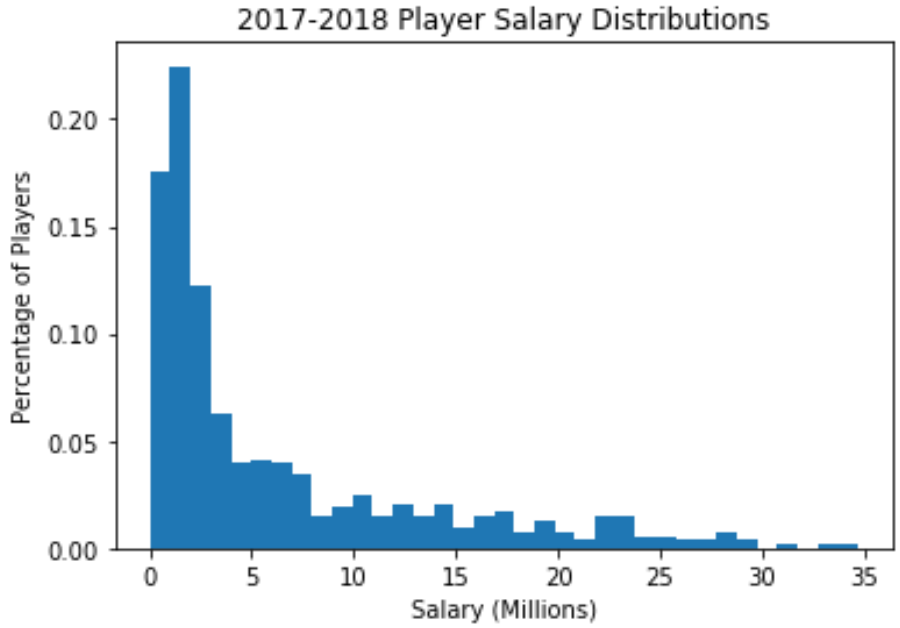

Figure 1. Player Salary Distributions (2017-2018)

Additionally, the methods I used could be used as a framework for current NBA player statistics to get results basically in live time or used for any season. The ideology behind this analysis could also be useful within other sports leagues, although some things would need to be changed to match the idiosyncrasies within said sport.

Project Proposal

As briefly mentioned, the goal of this project is to find the most undervalued NBA players. The idea of the analysis is to create the best predictive model using player in-game performance statistics as features and player's salaries as the target variable. Additionally, the root mean square error (RMSE) and R-squared values are compared to determine the most optimized model. To do this, two datasets were combined and used; NBA Player Salary Dataset (2017-2018) and NBA Players stats since 1950, both retrieved from Kaggle.

Implementation

To achieve the intended results, there were several things which needed to be done. First, the datasets were not in the ideal format for the problem statement. To account for this, they individually had to be cleaned before merging. This included converting values into integers, dropping missing values, and removing unnecessary duplicates. The most notable part of the

cleaning process was converting player statistics from 'totals' to a 'per-game' basis to eliminate the effect of amount of games played due to injuries, load management etc. Without this conversion, the results would be heavily skewed toward players who simply played more games or minutes.

After the datasets were merged there were 55 total variables, therefore feature reduction was implemented to remove less important features. Through research and trial and error, linear regression was found to be the most effective model used to achieve the desired results. The data were split into testing and training groups and fit to the models. After this, cross-validation was performed on the models and several combinations of features were tested and evaluated by comparing the RMSE and R-squared metrics. The RMSE is the square root of the variance of the residuals, and it indicates the absolute fit of the model to the data or in other words, how close the observed data points are to the model's predicted values. R-squared on the other hand is a relative measure of fit[1].

Results

Lower RMSE values and higher R-squared values indicate a better fit, thus by using these metrics the best combination of features was selected and used on the final linear regression model. To achieve the results, residuals were created to rank the players by taking the cross-validated estimated salary based on player statistics and subtracting it from the player's actual salary. A positive residual resembles overvalued players, and a negative resembles undervalued ones. The residual values were then sorted from least to greatest, ranking all players from the 2017-2018 NBA season from most to least undervalued. Figure 2 represents the top 10 results of my findings.

Rank	Player
1	Karl-Anthony Towns
2	Nikola Jokic
3	Joel Embiid
4	Zach LaVine
5	Kristaps Porzingis
6	Devin Booker
7	Jabari Parker
8	Myles Turner
9	Andrew Wiggins
10	Clint Capela

Figure 2. Top 10 undervalued NBA players from 2017-2018

Conclusion

The results are relatively promising according to cross-references, as well as when considering a large percentage of the players that made the top of the list were still on rookie contracts (known for being low-cost contracts) in the 2017-2018 NBA season. Additionally, the rookie contract players which appear on the top 10 were some of the most promising young athletes looking forward. To improve this analysis further, it would make sense to remove or highlight all players on rookie contracts because these players will undoubtedly be worth much more once their rookie contracts are terminated, meaning they will not remain undervalued to such extent. This could be done by merging these results with a player dataset containing player contract details. While Figure 2 only shows the top 10 undervalued players, the analysis ranked every NBA player from the specified season. A by-product of this is that overvalued players were also revealed, which could be valuable information to an organization. As briefly noted earlier, the methodology used in this case study could be applied by organizations to improve decision making. Additionally, this analysis could be run on multiple seasons to plot progression of a player's value over time. These methods or variations could also be used by an organization in tandem with additional models to improve the success and longevity of any sports organization.

Grace-Martin, K. (n.d.) *Assessing the fit of regression models*. The analysis factor. https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/