Evaluating NBA Players' Value in the Free Agency

Tego Chang

11/27/2021

Summary

In this report, we aim to evaluate the value of NBA players in terms of the salary they could earn next season, based on their performance in the current season. During the analyzing process, we have conducted data wrangling, exploratory data analysis (EDA), and modeling by multiple linear regression with and without considering a hierarchical framework. With our proposed model, we have provided a prediction of players' salary in the 2021/22 season with a 77% Hit Rate. Moreover, we analyzed the overpaid and lowballed cases and identified the metrics we shall additionally consider to further explore the topic.

Introduction

For general managers or the front office in NBA teams, one of the major tasks is to decide what contracts their teams shall provide for the players they are interested in based on these players' values from and off the court. As the budget for each team is regulated by the league, called (Team) Salary Cap, pursuing the players they want with a reasonable or even economical cost has been more critical than ever.

The primary question we would like to answer in this analysis is what statistics or factors could be influential to our forecasting of players' salaries next season. Further, we would also pay attention to if and what are the potential interactions between variables in the collected dataset.

Data

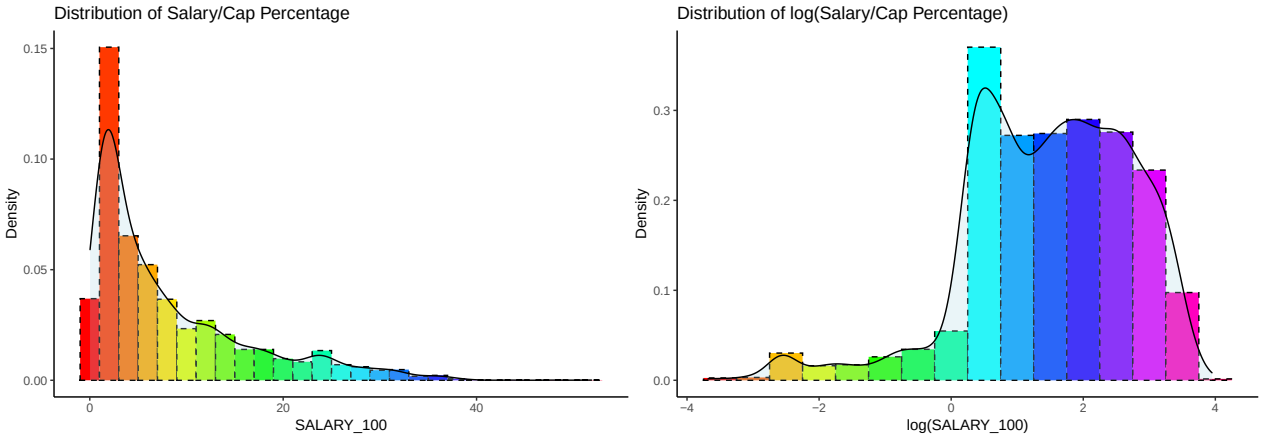
The response variable in this research is composed of two items, player's salary in the season, *SALARY*, and the salary cap for each team *SALARY_CAP*. These two types of data are collected from HOOPSHYPE and Basketball Reference websites; On the other hand, the response variable, *SALARY_100*, is obtained by diving *SALARY* with *SALARY_CAP* and multiple 100 to make it the percentage of the player's salary accounts for in the team's budget.

The predictors are separated into two categories, the traditional statistics and the advanced ones, and they are both collected from the official website of the NBA.The traditional statistics we picked were *AGE*, *GP* (game played), *W* (wins), *L* (Losses), *PTS* (average points per game), *REB* (average rebounds per game), *AST* (average assists per game), *TOV* (average turnovers per game), *STL* (average steals per game), *BLK* (average blocks per game), *3PM* (average three-pointers made per game), *+/-* (plus-minus, how many points the team can outscore its opponent when the player is on the court); The advanced ones are *OFFRTG*, *DEFRTG*, and *NETRTG* (correspondingly, the team's offensive, defensive, and overall performance scores when the player is on the court).

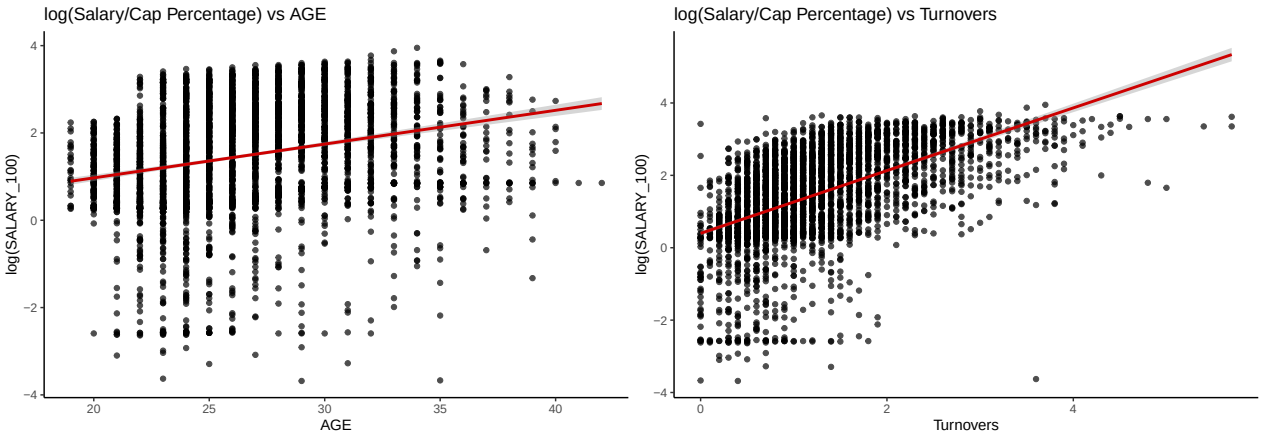
Once all the above are collected, we then perform data wrangling to make our dataset clean enough for the following analysis. Firstly, we merge the traditional and advanced statistics, and then we combine the player's salary with the player's performance. Lastly, we concatenate these observations on an annual basis. As the source data we collected were complete, we did not encounter any missing data issues. The arranged training data includes 3835 observations, which are the 10-year NBA players' data from season 2010/11 to 2019/20. The testing data is the one in season 2020/21. We plan to build a multiple linear regression model, train it

with the 10-season data, and apply the model to predict the player's salary in the 2020/21 season. Variables like *TEAM* and *SEASON* are considered as group variables when hierarchical modeling is conducted.

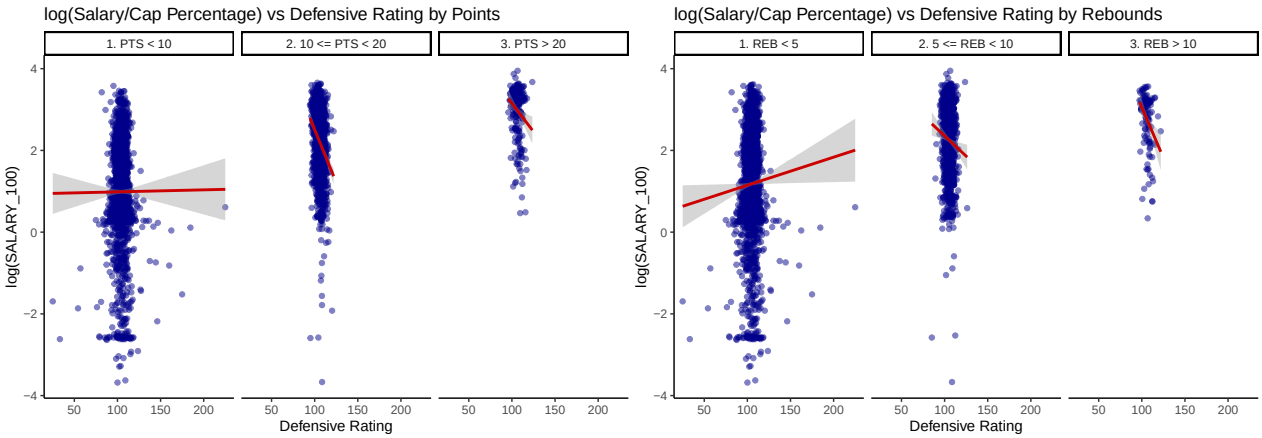
In the EDA stage, we first checked if our response variable, *SALARY_100*, has fitted normal distribution. We found it's not so we then have it transformed with logarithm to make it roughly follows the normal distribution assumption for the response in linear regression as the below figures show.



Then, we explored the relationship between predictors and the response. We observed some counter-intuitive trends for *AGE* v.s. *log(SALARY_100)* and *TOV* v.s. *log(SALARY_100)*. As we are talking about professional sports, we tend to think a younger player has the advantage to be faster and sustainable compared with the old ones, which might lead to better performance and higher salary; On the other side, having turnovers means you waste a chance to score, which is bad for the team, and it's reasonable to imagine it will have a negative relationship with the salary. However, it seems also not true based on the below figures. The phenomenon will both be explained in the following sections.



At the last in our EDA, we investigated if there are interactions between predictors. We found *DEFRTG* has interactions with *PTS* and *REB*. The latter one is again very counter-intuitive from the perspective of domain knowledge, basketball, as having more rebounds for a player tends to make people believe that the player performs well on the defensive end. We'll include these interactions during the model selection process.



Model

As the main purpose of this research is to make predictions, we tend to select fewer predictors which can more precisely stand for the player's value. Thus, we decided to go with stepwise BIC as our model selection approach.

Multiple Linear Regression To proceed with model selection by stepwise BIC, first, we need to decide the null and full models for the stepwise function. We prefer to include as few strong assumptions as possible, so we include no predictors in our null model. For the full model, we include all the main effects and the interactions we found interesting in the EDA, which are $DEFRTG : PTS$, $DEFRTG : REB$, and $DEFRTG : AST$. The outcome model's mathematical equation is as the following:

$$y_i = \beta_0 + \beta_1\ PTS_i + \beta_2\ REB_i + \beta_3\ TOV_i + \beta_4\ AGE_i + \beta_5\ GP_i + \beta_6\ W_i + \epsilon_i;$$
$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, \ldots, n.$$

where y_i is log of the player's salary versus salary cap in percentage (%).

We also noted that having logarithm transformation for the original response variable might not be enough for satisfying the normal assumption of linear regression. We also performed the boxcox transformation and obtained a lambda whose value is 0.25, which makes the y_i in the above formula be $SALARY\_100^{0.25}$.

For the proposed model above, we have conducted its assessment. First, as the below two figures, we confirmed that the constant variance, independence of errors, and normal distribution are met. Further, the linear relationships between response and all the predictors have all been verified. Thus, we confirm all the assumptions of linear regression are met.

To validate our proposed model, we applied RMSE as our evaluation metric. First, we made predictions for the testing dataset and the RMSE is 5.31. Then, we conducted K-fold validation to the training dataset with K = 10, and the average RMSE is 5.18. The results are considered acceptable as an initial model, but we do acknowledge there's still room to improve as when the salary cap increases, the exact error in the US dollar rises, too.

Hierarchical Multiple Linear Regression At last, we plan to consider the hierarchical framework and include group variables like *TEAM* and *SEASON* in the proposed model above. First, we need to decide whether we shall add both random slope and intercept or only random intercept is required. As during our EDA for the two group variables, we found that there's rarely trend change between the group variables and the other predictors, we concluded that we will simply consider random intercept for our model; Then, the second to decide is whether we shall add both group variables or only one. We performed an anova test for two potential models, the one with random intercepts on both group variables and the one with only

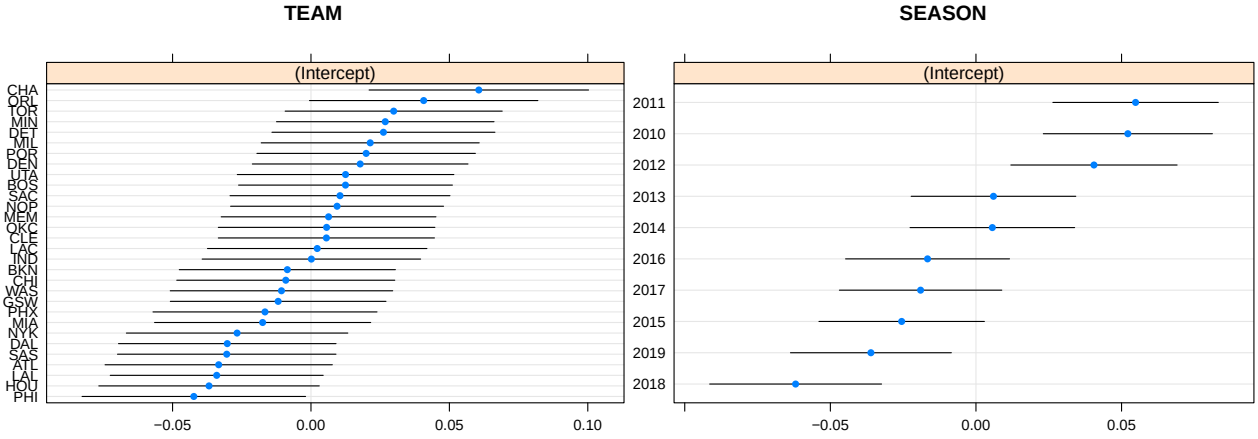
Table 1: Summary of the Proposed Hierarchical Model

	Dependent variable:
	SALARY_100^lambda_trans
AGE_c	0.0205^{***} (0.0183, 0.0227)
PTS_c	0.0317^{***} (0.0287, 0.0346)
REB_c	0.0364^{***} (0.0320, 0.0408)
TOV_c	0.0592^{***} (0.0390, 0.0794)
W_c	0.0052^{***} (0.0041, 0.0063)
GP_c	-0.0010^{***} (-0.0018, -0.0003)
Constant	1.5232^{***} (1.4936, 1.5528)
Observations	3,835
Log Likelihood	-663.2185
Akaike Inf. Crit.	1,346.4370
Bayesian Inf. Crit.	1,408.9560
Note:	*p<0.1; **p<0.05; ***p<0.01

	Groups	Name	Variance	Std.Dev.
1	TEAM	(Intercept)	0.00103	0.03206
2	SEASON	(Intercept)	0.00173	0.04158
3	Residual		0.08020	0.28319

Table 2: Random Effects of Hierarchical Model

the *TEAM* variable as a random intercept. The null hypothesis is rejected so we decided to apply random intercepts on both group variables. Our proposed model is summarized as below table:



For this model, we observed that the variance which could be explained by the two group variables, *TEAM* and *SEASON*, is only around 20% of the overall variance as the above table shows. This indicates the variance within groups has far from been fully explored and shows the topic deserves more investigations; On the other hand, the across-group variance among teams or seasons could be observed in the above two figures. Grouping by teams, Charlotte Hornets has the highest intercept while the Philadelphia 76ers has the lowest. Only these two teams have intercepts whose confidence intervals do not include zero, which indicates statistically significance. It could be interpreted as the former tends to have fewer players signed with big contracts and high salaries, while the latter tends to have more compared with other teams in the league. When a team has fewer players signed with really high salaries, the average salary for all the players in the team shall increase, considering each team spends roughly the same on their rosters; On the other, grouping by season, we found that before season 2015/16, the league had a positive intercept while after that, the intercept turned into negative. We would interpret that might result from the salary cap increases rapidly

since then. Together with the fact that not every player resigns a new contract with a higher annual salary when the salary cap increase, we could reasonably imagine that the percentage of player salary versus salary cap is likely to drop.

Lastly, as the interpretation for our proposed hierarchical model would differ by the random intercept of *TEAM* and *SEASON*, we simply picked Golden State Warriors (GSW) in the season 2019/20 as an example. Our model predicts that for a 26.6-year-old player in Golden State Warrior in season 2019/20 who had 9.3 PPG, 4.0 RPG, 2.0 APG, won half, 28 of the total games he played, 56, his salary in season 2020/21 is expected to be around 5.1M USD, which is the salary cap, 109140000, multiplying the response, 4.67%. Predictors, except for *GP*, all have a positive correlation with the response variable.

Conclusions

Our final model confirms that *AGE* and *TOV* are indeed statistically significant predictors and both are positively correlated with the response. We consider that could be relevant to the minimum salary guaranteed in the league increases as players play more years. In that way, players' age could roughly be treated as their years of experience in the league. Thus, it makes sense that our response will increase with *AGE*; As for *TOV*, generally, a player who handles the ball more often will likely to conduct more turnovers. In addition, usually, a team will try to have its best player to make plays for other teammates. Thus, no wonder higher turnovers seem lead to a higher salary.

On evaluating our proposed model's performance, we have further defined a metric called Hit Rate. We consider our prediction "hits" the real situation when the absolute difference between the predicted response and the truth is not greater than 5%, which is around 5.5 million USD. In such definition, our model has achieved a 77% hit rate to the testing dataset, the salary in the 2020/21 season. However, when we looked into some of the imprecise prediction cases as below table, which is filtered by the absolute difference between the predicted response and the truth is greater than 15%, we observed the potential limitation of our model and also the future work.

	PLAYER	Truth	Predicted	Predicted_2.5	Predicted_97.5
3	John Wall	39.42	14.70	13.65	15.81
16	Kemba Walker	31.11	13.47	12.82	14.15
19	Ben Simmons	29.36	12.81	11.93	13.75
22	Blake Griffin	28.83	8.01	7.71	8.32
27	Andrew Wiggins	28.09	12.25	11.78	12.73
29	Kevin Love	27.81	9.45	8.90	10.04
32	D'Angelo Russell	26.70	9.47	8.92	10.04
51	Gary Harris	18.62	3.43	3.28	3.58
116	Luka Doncic	9.05	30.42	28.49	32.45

Table 3: Underpaid and Overpaid Cases

Firstly, except Luka Doncic, all of the other players mispredicted by our model are overpaid. These players were either all-stars or highly potential when they signed a contract with their teams. As all of their contracts are on a multiple-year basis, they could still get the same amount of salary even when they played worse or even got injured during their contract. Thus, we might need to consider including the contract length as part of our response. Further, some predictors showing either a player's competitive spirit or healthiness could also be our potential predictors; Second, for the overpaid players, we shall try to clarify if there are some variables representing the player's potential or commercial value that our model oversees so that some of the general managers in the NBA are willing to provide a contract that our model identified as not worth it; Lastly, for lowballed players like Luka Doncic, most of them were still in their rookie contract, which limits the maximum salary they could earn in the early stage of their career. Thus, this implies that we shall also include contract type, e.g. Rookie, Bird, RFA, or more types, as one of the predictors.