ECO481: NBA Contract Prediction Using Machine Learning

Joonbum Yang [yangjoo5, joon.yang@mail.utoronto.ca]

Mai Dang Khoi Nguyen [nguy2106, bill.nguyen@mail.utoronto.ca]

En-Chan Sing [singench, jimmy.sing@mail.utoronto.ca]

December 15, 2022

1 Introduction

In 2016, the Golden State Warriors signed a record-breaking 54.3 million dollars 2 years contract with Kevin Durant, a 27 years old MVP and four times NBA scoring champion [1]. For an almost 24 billion dollars industry, the importance of fair monetary valuations in player contracts cannot be overlooked[2]. The cost of badly overestimated player contract does not reside only in its overvaluation but the team's performance in its entire season. However, the efforts to accurately predict a player's worth have been limiting. This paper will attempt to predict NBA contracts solely using its player statistics from the previous year and distinguish whether these players are being over/undervalued by using machine learning techniques.

Using machine learning techniques will allow us to take advantage of its flexibility in non-parametric modeling as past studies in sports analytics have been done mostly using linear regression. Ultimately, using machine learning to give accurate predictions on players' valuation can contribute to the reduction in market discrepancies in the NBA contract market.

2 Literature Reviews

There have been roughly ten academic papers covering a similar topic; however, each went in a different direction or had shortcomings. In 2014, Nuoya Li explored the relationship between players' last two years' statistics on their contract renewals and the terms of players' new contracts. However, this paper neglected advanced basketball metrics, potential indicators of players' performances that regular box-score does not cover.

Emirhan Ozbalta, Mucahit Yavuz and Tolga Kaya used machine learning methods, such as random forest and decision trees, to explore the relationship between basic bas-

[1] https://www.si.com/nba/2016/07/04/kevin-durant-free-agent-contract-signs-warriors-announcement

[2] https://www.cnbc.com/2021/03/22/nba-is-next-up-for-a-big-rights-increase-and-75-billion-is-the-price.html

ketball statistics, the video game NBA 2K20 player ratings and players' future earnings (. Their paper is potentially flawed for multiple reasons, most notably the use of players' video games, which are subjective to game developers' perspective, hence quite biased data. In their works, Ioanna Papadaki and Michail Tsagris showed that linear regression is not a good model choice as the relationship between basketball statistics and contracts is not linear (2020). Hence, techniques such as random forest or lasso regressions are better options.

3 Data and Methodology

The data set consists of 1389 players who were elected free agency between 2016 and 2022, excluding those who exercised player option or team option. The predictor variables include traditional box scores collected from Basketball-Reference by github user sumitrodatta and two advanced metrics from FiveThirtyEight: RAPTOR and WAR. RAPTOR complements box scores by taking advantage of player tracking data that better reflects how teams evaluate players. WAR, wins above replacement player, which benchmarks against a theoretical player who is on the fringes of NBA, provides an estimate of a player's value in terms of wins. The outcome variable, first year salary of a contract, is converted into percentage of that year's salary cap to account for yearly changes in the salary cap. Salary data are also gathered from Basketball-Reference.

As mentioned in Papadaki and Tsagris' paper, linear regression models are not ideal for portraying the relationship between the interested variables, so we have opted for tree-based regression methods to build our models. Before building our models, we randomly shuffled our dataset and split it into training and testing sets in an 80:20 ratio. Our first model is a regression tree, which recursively splits on predictors that will gain the most information leading to the variable of interest. Instead of using a single tree, the random forest technique takes advantage of bootstrapping within training data and produces aggregated predictions of variables of interest. The third model, gradient boosting, is another ensemble method which involves building a more robust model upon previously weaker-performing models to minimize its loss function and increase the predictive power. Lastly, extreme gradient boosting (XGBoost) is a technique which follows the principle of gradient boosting but with more regularized model formalization to control the risk of overfitting.

We then used different metrics to compare the performances of our four models. A higher R-squared value would indicate that the variables explain more variances for the variable of interest in our model. Another measure for quality of fit, mean squared errors (MSE), takes the average squared distance between the real data and predicted data in our testing set, which better models would have higher scores. Lastly, 5-fold cross-validation allows us to resample the data into different portions in five iterations and see how our model performs on average.

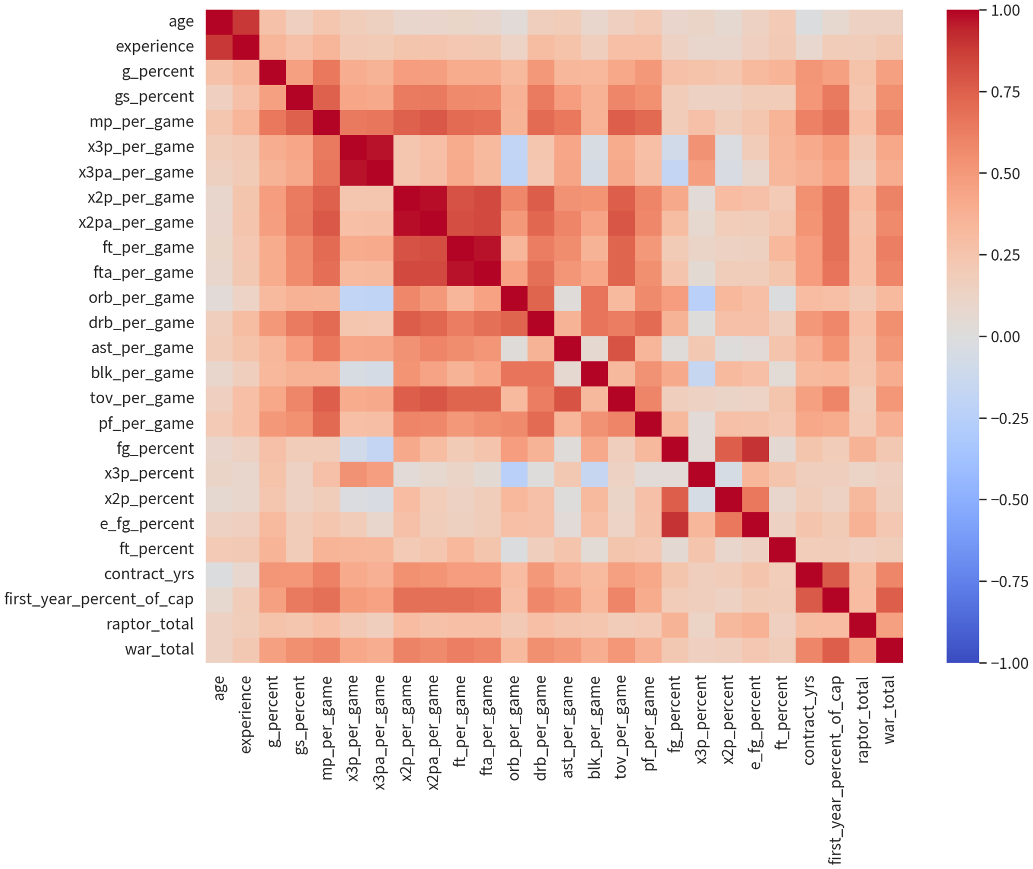

Figure 1: Correlation between interest variables

4 Results

The best performing model out of the four is the random forest model, which has the highest R^2, cross validation mean, and the lowest Mean Squared Error/Mean Absolute Error, followed by Extreme Gradient Boosting (Table 1). Minutes played per game, the most important feature of the random forest model, has nearly three times the effect of WAR, the second most important feature. This result is to be expected given more play time per game is highly correlated with higher box scores, which translate to higher first year salary according to our model.

In terms of overall market inefficiency, the total mismatch from 2016 to 2022 is predicted to be approximately 1390% of a team's salary cap, or roughly 1.71 Billion dollars using the current salary cap. At the individual player level, teams overpaid the most

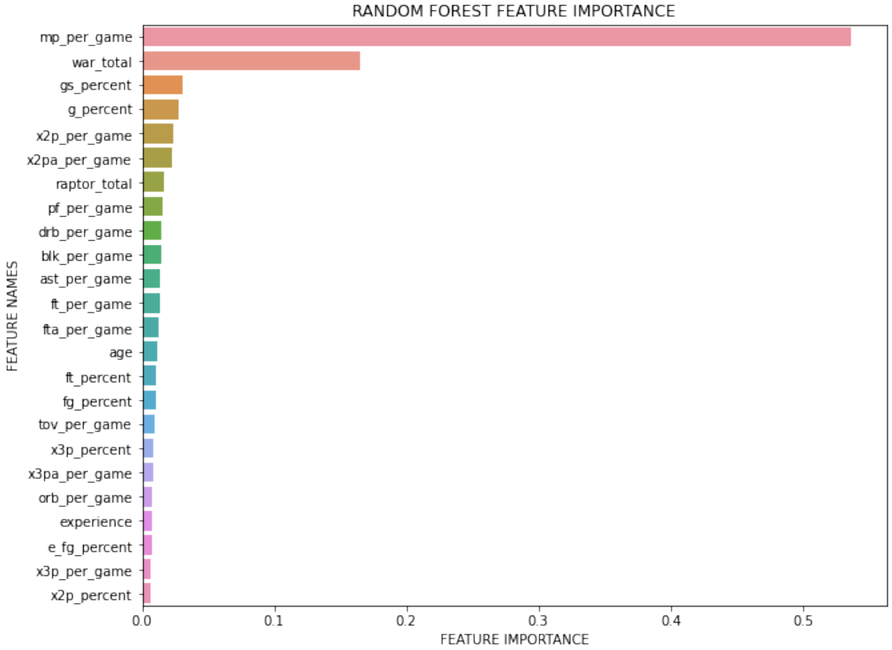

Figure 2: Feature importance of the random forest model

overvalued players of the last 6 years by roughly 9-13% (Table 3). The most undervalued players were underpaid by a similar degree of about 9-11% (Table 4).

Study has shown that star power has positive impact on consumer demand and team performance(Lawson et al., 2018, Lewis et al., 2018), thus, we also wanted to examine how big of a premium teams are paying to recruit popular players for the potential benefits. Surprisingly, based on our model the salary mismatch of popular players are rather small relative to their salary at around 1-2% (Table 5).

5 Conclusion

Using decision trees allows us to give predictions in player valuations while free from the limitations of parametric modeling such as requirements of parametric families and probability distributions. Utilizing these predictions can contribute to the reduction of market discrepancies, and asymmetrical information, in contract markets as players can be undervalued as much as overvalued.

5.1 Drawbacks

One of the most detrimental drawbacks of this method is the lack of avoidance of Endogeneity in variables. Player stats are highly likely to be influenced by variables within the model such as average play time per game. It is also not liberal of the existence of confounding variables such as past contract values.

This paper intentionally excluded making predictions based on past contract values due to two reasons. First, it was to minimize the effect of each player's name value which the salary will likely portray. Second, it came from the belief that wages are sticky and will likely increase over the year despite players' current stand on performance especially since contracts are usually settled for more than a year. However, previous contract values could encompass other information that the teams used for their valuation whilst that information is not available to the public.

6 Figures

Table 1: List of Variables

Variable	**Description**
first_year_percent_of_cap	first year salary in a player's contract as a percentage of team's salary cap
age	Player's age
experience	Years of experience in the NBA
g_percent	Percent of game appearances in the season
gs_percent	Percent of games started in the season
mp_per_games	Minutes played per game
x3p_per_game	Expected three point shots made per game
x3pa_per_game	Expected three point attempts made per game
x2p_per_game	Expected two point shots made per game
x2pa_per_game	Expected two point attempts made per game
ft_per_game	Free throws made per game
fta_per_game	Free throw attempts per game
orb_per_game	Offensive rebounds per game
drb_per_game	Defensive rebounds per game
ast_per_game	Assists per game
blk_per_game	Blocks per game
tov_per_game	Turnovers per game
pf_per_game	Personal fouls per game
fg_percent	Field goal percentage
x3p_percent	Expected three point shot field goal percentage
x2p_percent	Expected two point shot field goal percentage
x_fg_percent	Effective field goal percentage
ft_percent	Free throw field goal percentage
contract_yrs	Length of contract (non-guaranteed contracts are assigned 0 years)
raptor_total	Player's RAPTOR statistic
war_total	Wins above replacement player

Table 2: Comparison of Model Performance

Predictor	R-Squared	Mean Absolute Error	Mean Squared Error	CVS Mean	CVS Standard Deviation
Decision Tree	0.5069	0.0244	0.001945	0.541496	0.041196
Random Forest	0.7956	0.0179	0.000806	0.734524	0.020806
Gradient Boosting	0.7773	0.0183	0.000878	0.715102	0.024442
XG Boosting	0.7908	0.0179	0.000825	0.732674	0.023368

Table 3: Top 5 Overpaid Players Between 2016-2022

Player Name	Season	Age	Experience	Actual 1st year % of salary cap	Predicted 1st year % of salary cap	Difference
Timofey Mozgov	2016	29	6	0.1700	0.0401	0.1299
Zach Lavine	2022	26	8	0.3300	0.2302	0.0998
Dirk Nowitzki	2016	37	18	0.2656	0.1672	0.0984
Bradley Beal	2022	28	10	0.3850	0.2888	0.0962
Klay Thompson	2019	28	8	0.3000	0.2153	0.0847

Table 4: Top 5 Underpaid Players Between 2016-2022

Player Name	Season	Age	Experience	Actual 1st year % of salary cap	Predicted 1st year % of salary cap	Difference
Carmelo Anthony	2018	33	15	0.0235	0.1526	-0.1291
Victor Oladipo	2021	28	8	0.0213	0.1468	-0.1255
Ish Smith	2016	27	6	0.0637	0.1570	-0.0933
Derrick Rose	2017	28	8	0.0214	0.1126	-0.0912
Miles Bridges	2022	23	4	0.0000	0.0888	-0.0888

Table 5: Popular Players Contract Predicted Actual

Player Name	Season	Age	Experience	Actual 1st year % of salary cap	Predicted 1st year % of salary cap	Difference
Kawhi Leonard	2019	27	8	0.3000	0.2999	0.0001
Kevin Durant	2016	27	9	0.2819	0.2897	-0.0078
Lebron James	2018	33	15	0.3500	0.3284	0.0216
Nikola Jokic	2018	22	3	0.2458	0.2495	-0.0037
Stephen Curry	2017	28	8	0.3500	0.3309	0.0191