The Effect of Salary Allocation on Team Success in the NBA

By: Bradley Brown, Andrew Cannon, Isaac Corcoran, Chris Harden,
Felicia Seng, Jordan Khamvongsouk

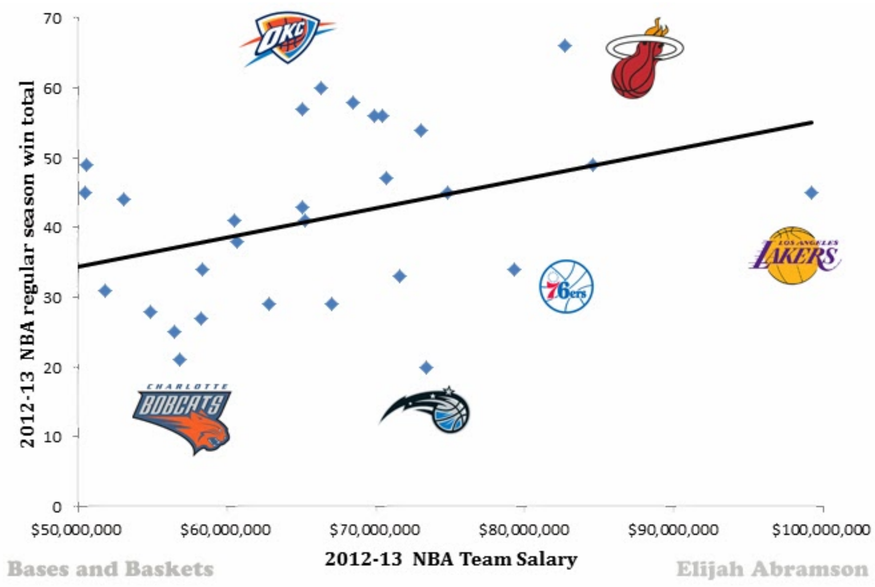
The world of sports has come a long way since it's beginning. Players are learning to improve and refine their game and skills every day. Today, coaches and staff members are becoming more and more equipped with the latest technology to help improve their team and organization. Sports analytics is changing the world of sports as we know it, pioneering a data-first mentality when assessing players' and teams' quality. This relates not only to helping players increase their physical and mental performance, but also with business and recruiting decisions as well. By leveraging data that is already being collected or by going out and collecting the data themselves, sports analysts are utilizing the process of the scientific method alongside statistical modeling to give players and organizations a competitive advantage over others on and off the court.

In our research, we wanted to explore the relationship between players' salaries and their teams' performance in the NBA. We wanted to see how allocation of salary cap correlates to the success of the team. We will discuss the background and supporting research from others, our assumptions for the research, our own hypothesis, and finally our approach to analyze the data and to create a model to predict overall NBA performance. One example that motivated this research stemmed from the Los Angeles Lakers in the 2010's. The Lakers had an injured and aging superstar during this time, Kobe Bryant, who they paid like he wasn't injured or aging. Throughout the decade, even after his retirement, the Lakers generally underperformed compared to expectations. We wanted to research if this could have had anything to do with the way salary was allocated within the team.

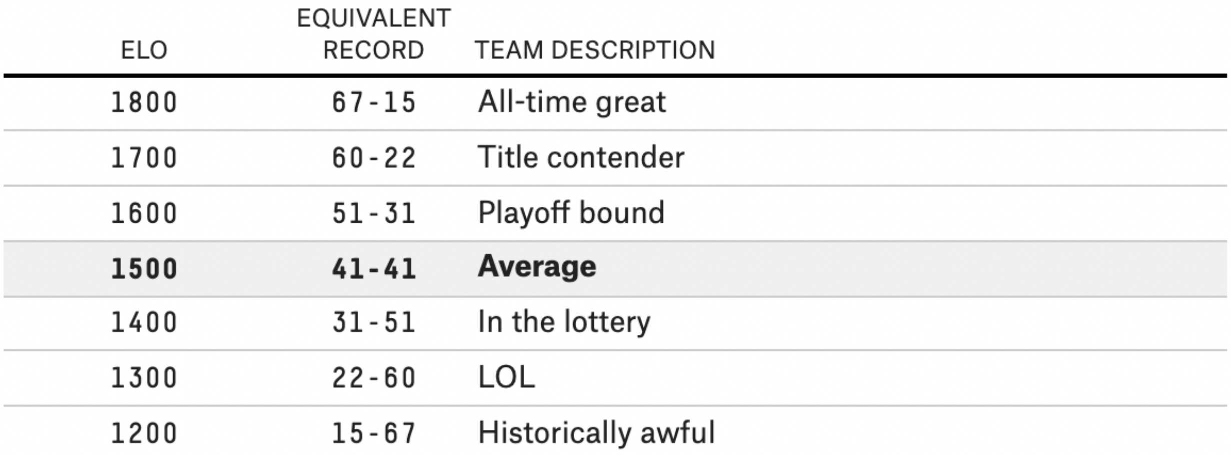
Similar to most aspects in business and in life, people invest in things that add value and contribute to their goals. In the NBA, a player that performs well (over a period of time) is usually compensated with a sizable contract and gets paid more than one who has been poor (or not performing) historically, or is new to the league in general. So it is logical that a team which has the best players in the league will cost the owners more money. However, the best players in one season are not always the best players in the next season and it is important to get contributions from players on the team who aren't expensive superstars. This is where our research is aimed, to highlight the patterns and trends in contracts and salaries for NBA teams and compare them to how those teams performed. One of our goals is to highlight teams who were able to get more on the court than what they paid for when compared to their other seasons and the other teams in the league.

There are many moving parts we are dealing with in our analysis, unfortunately it's never quite as simple as "pay more money and then you'll win more games". According to research done by SangYeon Choi titled "*Will the NBA Player's salary contribute to the team's Win?*", there is a strong positive correlation between minutes played and player salary. And on the surface that makes sense, but we need to do more than take this at face value. Players who play well are going to play more minutes but that does not mean anyone who is good is playing a lot of minutes. Injuries often prevent some of the best players from contributing to wins. But these players are still worth their salary to be a part of their team.

There has been research done like Choi's on how salary is correlated to winning in the past, but it has mainly been focused on how values such as total salary or max salary match up with teams' win totals. The NBA has also had a lot less of this kind of study than other sports leagues, such as the MLB. In general, there has been a positive correlation between salary and winning, as most of the best players in sports are paid large salaries and contribute the most to winning. But often a team with a very large win total or one that goes deep into the postseason will turn out to have a low total salary and no max players. This comes from teams who have cheap superstars, who may be on rookie deals or took discounts. It also comes from finding valuable role players for cheap. One case study that was done on the 2012-13 season in the NBA produced this chart, illustrating a small example of the relationship between winning and salary.



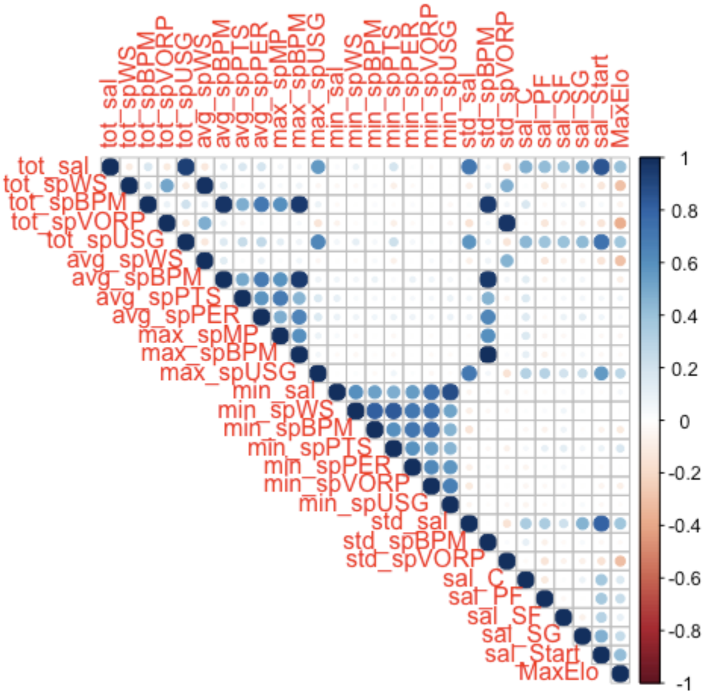
There has also been a lot of research done on whether or not player’s in the NBA and other sports leagues get paid enough based on how much revenue the sport brings in and how important the players, especially the best ones, are to that revenue. This is insightful to our project, as you could look at superstars as being on a discount since their salaries are maxed out. Ultimately in our research, we are comparing teams to each other within the NBA. We have combined this idea of salary and winning with advanced statistics such as Win Shares, Box Plus Minus and VORP among others that measure the statistical impact of players in the NBA. Our goal was to measure how teams allotted their salary based on players’ performance and how that contributed to their performance on the court as a team. We researched multiple different representations of team success both in the regular season and postseason and ultimately decided to use a measure created by Five-Thirty Eight, ELO rating. ELO rating changes throughout the season after each game, based on the result. We chose to use the max ELO that a team had throughout the season as this was a good representation of the best the team could be during the season with the salary as constructed. Elo can be interpreted as follows:



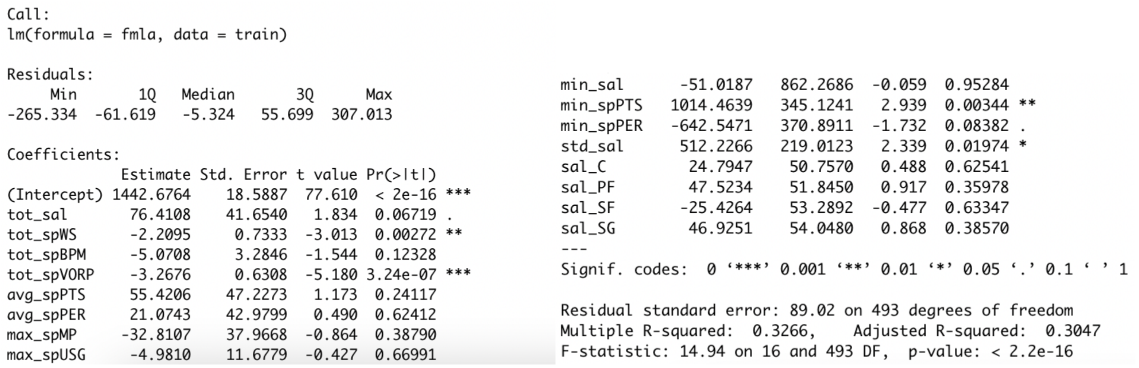
ELO	EQUIVALENT RECORD	TEAM DESCRIPTION
1800	67-15	All-time great
1700	60-22	Title contender
1600	51-31	Playoff bound
1500	**41-41**	**Average**
1400	31-51	In the lottery
1300	22-60	LOL
1200	15-67	Historically awful

In order to properly analyze our data, we must point out our assumptions and hypotheses. We believe teams that efficiently use their allotment of the salary cap will have an overall improved team performance and will be more likely to reach the playoffs. We must also point out potential flaws in our analysis and the assumptions we made to avoid those flaws. Like most professional sports, the NBA deals with many variables when it comes to how much its players actually play. For example, when a player is injured, he doesn’t necessarily contribute to the overall success of the team, but he does still contribute to the team’s salary cap usage. For example, in 2013, Kobe Bryant tore his Achilles and proceeded to play only 6 games the next season while the Lakers still paid him $30 million dollars. The team’s max ELO dropped from its typical 1600+ value to a lower 1500 point value without him. This is a byproduct of injuries in the NBA and there currently isn’t a good way to control for this. Another assumption we are making is that the end of season max ELO we have is not affected by mid-season trades, or in other words a team would not have a higher or lower max ELO if a player they acquired mid-season was with the team the whole season. A last area of concern is individual player motivation to perform throughout a season. Player attitude cannot be controlled, so we are assuming they are motivated to perform their best every game. With these assumptions in mind, we attempted to model our data using both a random forest model and a linear model.

Given the variables, we are seeking to build a model where we can predict the max Elo of any team given a set of salary-related variables within a certain range of that team's actual max Elo. In order to create our models and test them, we randomly split our data set into training and testing data points, with 80% of the data being in training and 20% being in testing. Because our original dataset suffered from multicollinearity, we used the lasso method to select our variables for use in the final model in order to enhance the prediction accuracy and interpretability of the resulting statistical model. Through lasso, we narrowed down our variables in use from the original 48 to 27. With the information given, we were able to select the variables through lasso and create a new master dataset to predict the most accurate ELO rating.



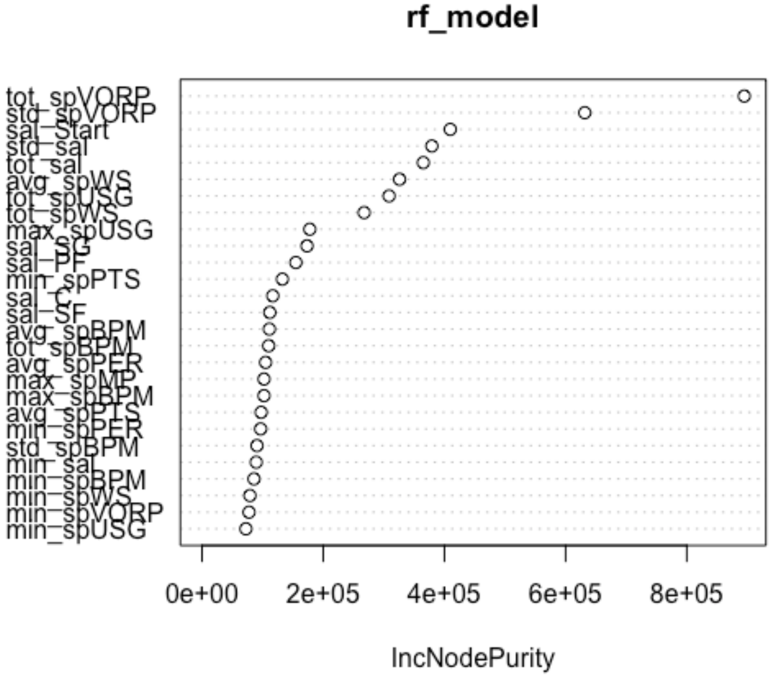
Since there was still multicollinearity present in the model, we removed as best we could any variables that were obviously heavily correlated with one another. One example of above is max_spBPM, which is almost a correlation of 1 with both tot_spBPM and avg_spBPM. This process left us with a final set of 16 variables to use for our model. We ran the linear model using the independent variables chosen through variable selection and the dependent variable of maximum Elo rating. Below is a summary of the linear model.



```
Call:
lm(formula = fmla, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-265.334  -61.619   -5.324   55.699  307.013

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 1442.6764    18.5887  77.610  < 2e-16 ***
tot_sal       76.4108    41.6540   1.834  0.06719 .
tot_spWS      -2.2095     0.7333  -3.013  0.00272 **
tot_spBPM     -5.0708     3.2846  -1.544  0.12328
tot_spVORP    -3.2676     0.6308  -5.180 3.24e-07 ***
avg_spPTS     55.4206    47.2273   1.173  0.24117
avg_spPER     21.0743    42.9799   0.490  0.62412
max_spMP     -32.8107    37.9668  -0.864  0.38790
max_spUSG     -4.9810    11.6779  -0.427  0.66991
min_sal      -51.0187   862.2686  -0.059  0.95284
min_spPTS   1014.4639   345.1241   2.939  0.00344 **
min_spPER   -642.5471   370.8911  -1.732  0.08382 .
std_sal      512.2266   219.0123   2.339  0.01974 *
sal_C         24.7947    50.7570   0.488  0.62541
sal_PF        47.5234    51.8450   0.917  0.35978
sal_SF       -25.4264    53.2892  -0.477  0.63347
sal_SG        46.9251    54.0480   0.868  0.38570
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 89.02 on 493 degrees of freedom
Multiple R-squared:  0.3266,    Adjusted R-squared:  0.3047
F-statistic: 14.94 on 16 and 493 DF,  p-value: < 2.2e-16
```

After creating the model using the 80% set aside for training. We tested the model using the other 20% of our data. We checked whether the predicted max Elo rating was within the same range (using the breakdown on page 4) as the actual max Elo rating or not. By this measure, the linear model got the correct range for the team 73.44% of the time. The biggest difference between predicted and actual for the linear model was the 2009-10 New York Knicks, and the model was 259 points off, predicting the Knicks to be a historically awful team with a max Elo of 1215, when in reality they capped out just under average at 1474, but still not making the playoffs.

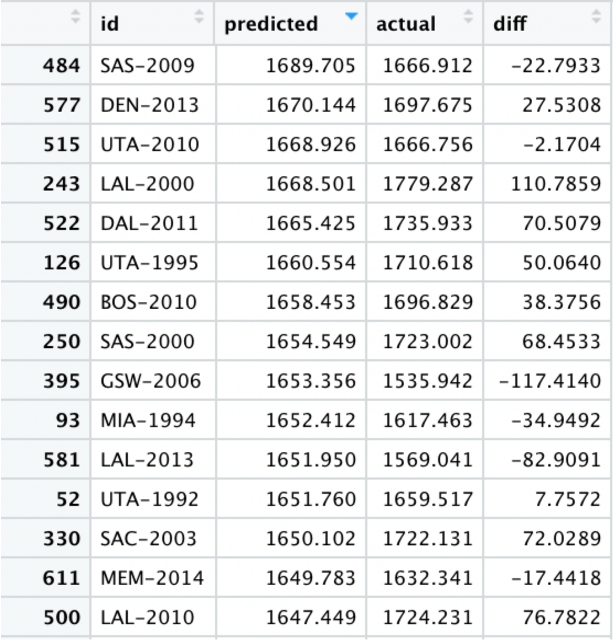
The second model we created using our data was a random forest so we could compare its performance with our linear model. Similarly to the setup of our linear model, we used the Lasso variable selection method and then inputted the chosen variables into our random forest model. Here is a variable importance plot from the model:



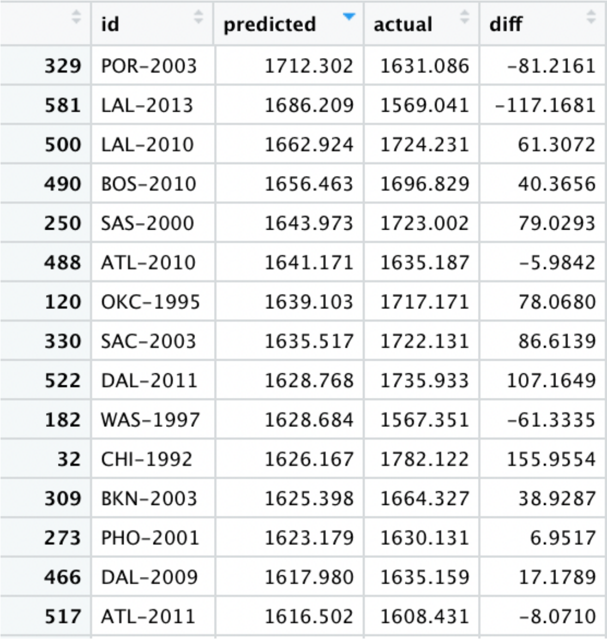
Again, with a similar method to the linear model, we trained using a randomly selected set of 80% of the data and then tested the model by predicting the other 20%. We then measured the effectiveness of the prediction based on how many of the max Elo estimates were in the same range (based on the ELO interpretation chart above) as the actual max Elo for the team that year.

With an accuracy of 85.94%, the random forest model performed better than the linear model. In the results from the random forest model, the worst estimate was the 2003-04 San Antonio Spurs, predicting them to be an average ball club (1551), when in reality they capped out as a legitimate title contender with a max Elo of 1764, sweeping the Grizzlies in the first round and getting up 2-0 on the Los Angeles Lakers before losing the next 4.

When comparing the two models, we ultimately chose to use the random forest model as our final model. Although the linear model is slightly easier to interpret, the random forest can give us much more accurate predictions, while still providing us with an understanding of which variables are the most important to the final result of the model. Our model and results are quite heavy statistically and require a good amount of unraveling to interpret to a team or coach, so using the random forest seemed to make the most sense. It adds accuracy and precision, and doesn't add too much confusion relative to the complicated nature of our final model and presentation. Here is a snapshot of some of the results in a table, random forest on the left, linear model on the right:



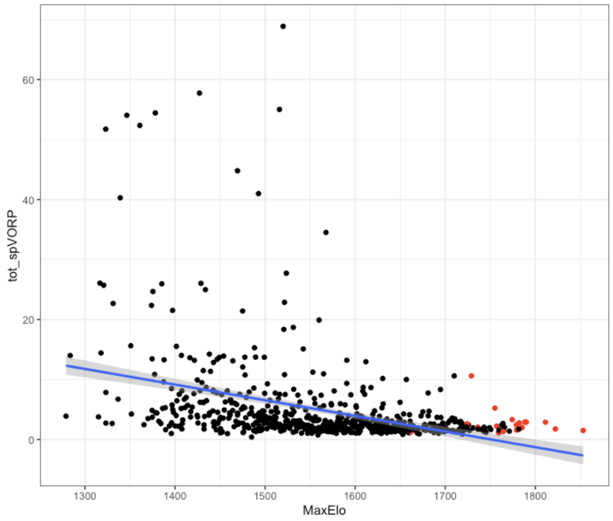
	id	predicted	actual	diff
484	SAS–2009	1689.705	1666.912	–22.7933
577	DEN–2013	1670.144	1697.675	27.5308
515	UTA–2010	1668.926	1666.756	–2.1704
243	LAL–2000	1668.501	1779.287	110.7859
522	DAL–2011	1665.425	1735.933	70.5079
126	UTA–1995	1660.554	1710.618	50.0640
490	BOS–2010	1658.453	1696.829	38.3756
250	SAS–2000	1654.549	1723.002	68.4533
395	GSW–2006	1653.356	1535.942	–117.4140
93	MIA–1994	1652.412	1617.463	–34.9492
581	LAL–2013	1651.950	1569.041	–82.9091
52	UTA–1992	1651.760	1659.517	7.7572
330	SAC–2003	1650.102	1722.131	72.0289
611	MEM–2014	1649.783	1632.341	–17.4418
500	LAL–2010	1647.449	1724.231	76.7822

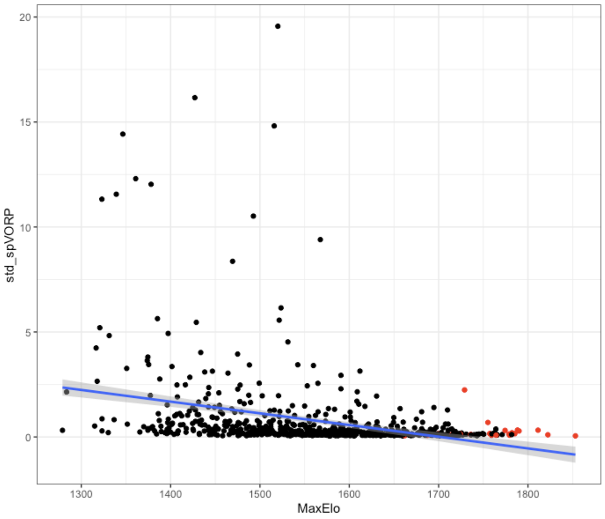


	id	predicted	actual	diff
329	POR–2003	1712.302	1631.086	–81.2161
581	LAL–2013	1686.209	1569.041	–117.1681
500	LAL–2010	1662.924	1724.231	61.3072
490	BOS–2010	1656.463	1696.829	40.3656
250	SAS–2000	1643.973	1723.002	79.0293
488	ATL–2010	1641.171	1635.187	–5.9842
120	OKC–1995	1639.103	1717.171	78.0680
330	SAC–2003	1635.517	1722.131	86.6139
522	DAL–2011	1628.768	1735.933	107.1649
182	WAS–1997	1628.684	1567.351	–61.3335
32	CHI–1992	1626.167	1782.122	155.9554
309	BKN–2003	1625.398	1664.327	38.9287
273	PHO–2001	1623.179	1630.131	6.9517
466	DAL–2009	1617.980	1635.159	17.1789
517	ATL–2011	1616.502	1608.431	–8.0710

The question after viewing so much data is always the same. So what? It can be tough in a world of modern statistics to wade through the never-ending numbers, especially when basic statistics like points per game or rebounds per game are sometimes easier to understand, but with our advanced statistics intertwined with salary data, we can quickly and accurately tell which teams are most likely to perform well that season, make the playoffs and be title contenders. It can also let your front office know about how much you should be paying players based on their statistics in order to find more success as a team. To be clear, our model doesn't necessarily predict NBA champions, but it will be enough to help create a playoff caliber team which is the first step towards a championship and for many is considered to be a good season.

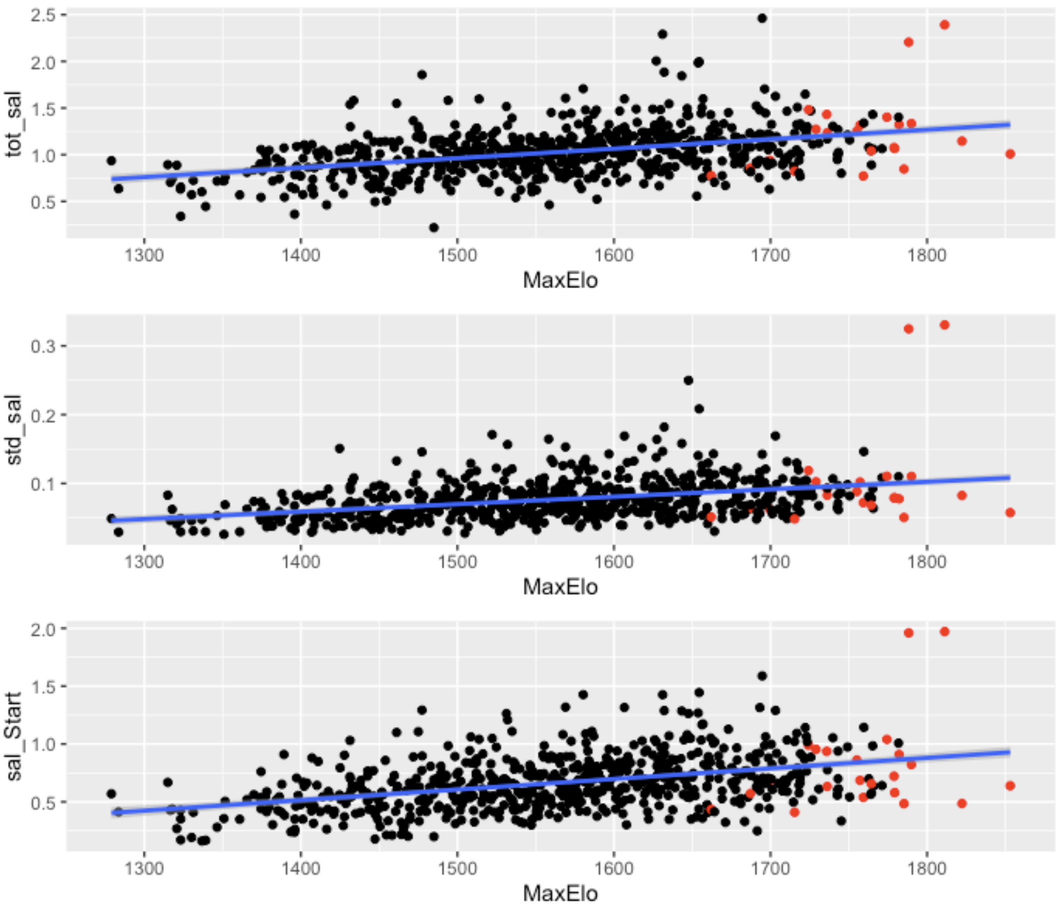
Our model could have an important impact on the way building an NBA team is managed. Our research suggests that as a team spends less of its salary cap for their player production, especially measured by VORP, they tend to perform better during the season, specifically having a higher maximum ELO rating. The total percent of the cap that was spent per 1 unit of VORP (tot_spVORP) was the most predictive variable in our data set. This suggests that a good strategy to take before signing players would be to figure out a reasonable goal for this value for your team based on existing contracts and expected extensions. Then sign and/or draft players based on their VORP from the previous season and how it will carry over to the next. The application of this research could lead team managers to more precisely decide how to allocate their salary to new players and how to increase their overall success.





Above (champions in red) is the relationship between tot_spVORP and max Elo rating on the left. As can be seen, paying a higher portion of the cap for VORP is not very indicative of a good team. Interestingly, it can be seen on the right that the standard deviation of this same value, which is also important to our model, has a similar relationship. If teams are successful at paying players consistently for the VORP they produce, they can be extremely successful.

Another important takeaway comes from the next few important values in our model, sal_Start (proportion of the salary cap allocated to starters), tot_sal (total proportion of the salary cap used) and std_sal (the standard deviation of the proportions of the cap used by the salaries of the players contracts), all displayed below. As can be seen, these are positively correlated with max Elo. This indicates that having high salaried starters is important to winning, which makes sense considering the well-documented star power of the NBA. Arguably the only 3 teams to win championships in the last two decades with fewer than two top-10 players are the Detroit Pistons in ‘04, the Dallas Mavericks in ‘11, and the Toronto Raptors in ‘19. For front offices assembling a team, this means generally if the team is striving to be the best, they need to be ready to open the checkbook, but doing so doesn’t guarantee success.



In further studies, we believe we could make this analysis more useful by modifying our variables to take player injury into account. We would standardize all metrics by either number of games played or number of minutes played. Injury to major contributors on any given team can have a major effect on max ELO and team success during the season as we previously stated in our assumptions. A potential weakness in the predictive power of our model we have seen is the skewed results that could come from major team contributors becoming injured. If a player were injured and unable to perform for extended time, other players would naturally have to step up and take over the production in his place. Since these players would likely have a smaller percentage of salary cap contributed to them - not being starters - it could create a false prediction that since a smaller percentage of salary cap was being used per player production, the team would actually perform better. Since the loss of a starter or major contributor rarely if ever improves team production and success, this could weaken our model's predictive capability.

Something else that might help to incorporate in future studies would be more data from the more niche NBA Collective Bargaining Agreement rules. Since the NBA has a soft cap rather than a hard cap, there are numerous ways to spend more money than you are technically allotted. Studying things like whether or not using Bird rights or poison pills or other exceptions to exceed the salary cap is a good idea or not would probably be a good use of time. Bird rights are used all the time to re-sign their best players that they've kept for three or more years, and it allows teams to exceed the salary cap. Poison pills make trading good players with little experience a little trickier for teams to get a good deal salary-wise. Are these salary exceptions worth it to make a team better? It is also worth noting that the salary cap rules and exceptions are frequently changing, and although we accounted for inflation and changes in the cap in our calculations, we didn't necessarily account for these exceptions and more minute rules changing over time. It might help to focus our data on a smaller window of time where there wasn't much rule changing or salary cap change either for convenience.

Individual contracts may be created under the collective bargaining agreement. Maximum salary may change on the player depending on the number of years that a player has played basketball and the salary cap. An exception towards this is that a player may sign 105% of his previous contract even if the new contract is higher than the league limit. Under CBA, rookie scale salary assigns salary depending on draft position. Each NBA team is also eligible to nominate a designated player in which a player can go for a 5-year contract extension, instead of being held to the standard 4-year restriction. There are many other rules in the CBA that allow a difference between the salary pay of each player. With the CBA change during different time periods and its allowance for individual contracts in regards to salary cap, influencing variables may have a cause and effect on the prediction of success. Understanding this factor, further improvements may be made on the study of NBA salary allocation and team success.

Tied to the intricate details of the CBA is another potential avenue for future research. When doing our analysis, we found that older data (1990-2000) was not very predictive of newer data (2010-2015) when separated. We observed from trends in the data, that this mainly came from the way the handling of the salary cap has changed extensively over time. For example, the '97 and '98 Chicago Bulls paid Michael Jordan more than $30M, which at the time was more than 120% of the salary cap! Under the current rules, this is not allowed, so these kinds of actions by a front office could never happen. Situations like this make older salary data quite different from newer data, and less predictive of how salary can best be allotted in today's NBA. Future studies could benefit from splitting their research up by CBAs. We believe this would help identify clearer trends and filter out some of the noise.

Looking within the dataset, it is apparent that each player has different characteristics. Their age and experience are huge factors that contribute to and alter their chance of success. In further studies, we can take a deeper look into bench players. Bench players are generally paid a different amount of salary depending on their years of experience yet their contribution to the team's chance of success may not be as much for the season. For example, Jeremy Lin was mostly benched for the 2018-2019 NBA season. He played a total of one minute for his team, Toronto Raptors, during the NBA finals. Yet, his team was able to win the championship. During that one minute, he likely did not contribute a lot toward the team winning. Thus, a further study with a dataset separating the benched players and starters of each season could be informative. A new study focusing on the main players of the team for the season could be more effective.

Another potential improvement on our research that could be made would be the adjusting of the predictive variables we used. Most of our variables take some unraveling to understand, since we were focused on combining them for our statistical model, and not necessarily for descriptive purposes. For example, we currently have a variable in our model that is the total of all the players "spVORP" which represents the percent of the salary cap a player takes up in a given year divided by their percentile among all players that season in VORP. This is not very intuitive and there is certainly room in future analysis to improve these values into something more understandable to the average fan. These improvements may even improve the model overall.

Data analytics have an important role in modern sports, and the NBA is no exception. Analyzing how teams use salary to maximize their potential is pivotal to performing well in the league. We believe our analysis can lead to improved decision-making in allocating salary and can help teams gain an edge over their competition.