Alvaro Hu

SpringBoard

August 29, 2018

Capstone 1 Milestone Report

I. The Proposal

The purpose of my project is to give a salary, a finite and concrete number, to players in the NBA based solely upon their stats. I will also be working on figuring out which game statistics allow for not only the biggest increase in salary, but also the ones that are *most likely* to give a player an increase in their salary. This is useful for a wide variety of reasons, but mainly for these two; that General Managers (GMs) are informed with an actual number for how much a player *should* be worth based on their stats so that they can make decisions on players, bu also to give the players an idea on what to improve in their game in order to get as significant a pay increase as possible.

To do this, I went online onto basketball-reference.com and downloaded three datasets. One data set with all of the players' names with their salaries for the upcoming season, another with all of the players' names with their common per-game stats, and a third dataset to retrieve information on some more advanced statistics, like Win Shares, Player Efficiency Rating, and Assist Percentage. The reason why I used next year's salaries and last years statistics is because it is common practice to sign extensions and make deals over the offseason based on a player's previous season stats for the next year. I merged these three datasets together using the player's name as the axis, and got a data set with far too many features. I used the .drop() feature in pandas to drop a majority of them. I still kept quite a few features, but only to see if they would have an effect on the salary in a later stage of the project.

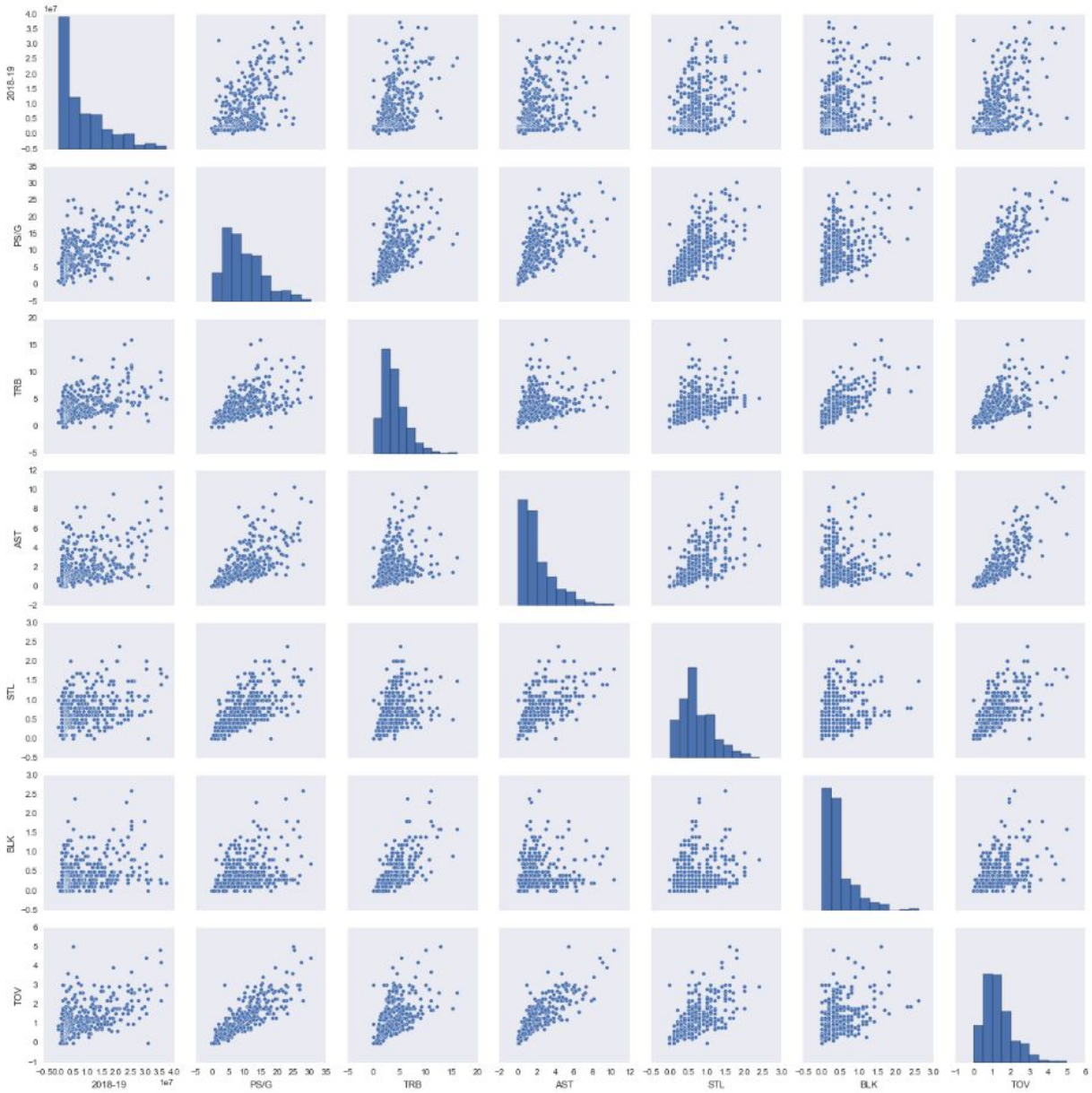
II. The Wrangling

When I first read the data into my project, there were a lot of odd things in the data that I had to deal with. For one, the 'Name' column had a string attached to it so that the player's name would be something like "Stephen Curry/stephencurry". So I had to split that column by the "/" figure before merging all of the data. There were also many players who showed up multiple times in the per-game dataset, since they played on multiple teams. To get around this, I checked to see how many players appeared with "TOT" in the Team column, which indicates that they played on multiple teams, and were given a Total column. The number I got was 59 players. I then counted up how many times these players showed up in the dataset, and found that these 59 names appeared 183 times, which makes sense because each of these players will be in at least 2 other rows. So I used the .drop_duplicates() attribute and kept the topmost column of the occurrences, since that column was the TOT row. I then merged the salary dataset with this one, but since the salary dataset included a row with the team that each player ended the season in, I was able to drop the Team column from the first dataset, throwing out all of the "TOT" entries.

From here on, the rest of the data wrangling was fairly straightforward. The only thing that I had to do was get rid of the "$" sign in the salary column and turn them all into integers, so that I can order them correctly. They were floats at first, which meant that the numbers were read in an alphanumeric fashion. For example, 10,000 would be placed ahead of 999,999 because "1" technically came before "9" in the alphabet. I also decided to drop player's who did not play last season, as it will be impossible to predict their salary from games they never played, and also player's without a salary for the upcoming year, as I will be unable to use them in my regression to predict a players salary. I then output my dataframe as a csv called 'Total2018.csv'.

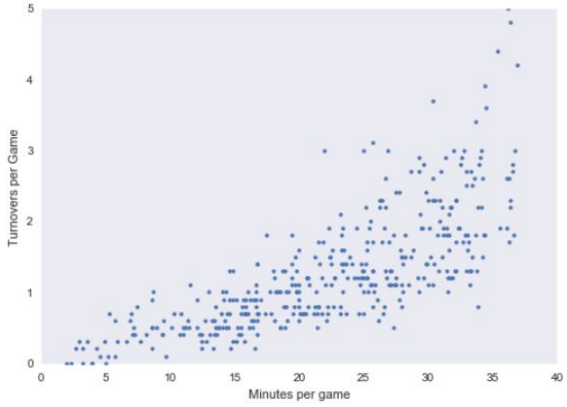
III. The Visual Exploration

After I wrangled all of the data and came up with a dataset with all of the features I wanted to use, I decided to visually explore the data to see if, right off the bat, I might be able to see some patterns. The first thing that I did was to make a few pair plots. I made one with the entire season stats, like games played, win shares, etc., and I made one with the per game stats. Below is the pairplor of the per-game stats.



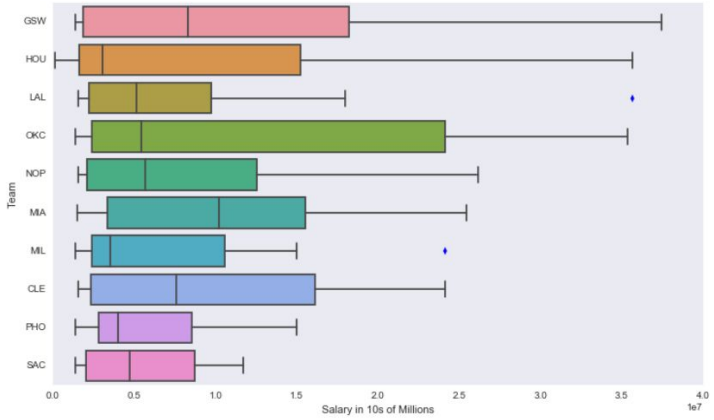
One interesting thing to take note of from this pair plot is that almost every statistical category has an almost linear relationship with every other statistical category. This may lead to a low number of principal components. The other thing to take note of is the non-normality of the distributions of these stats. The other pairplot shows that many of those stats are correlated too, but not quite as much. From these two pairplots, I found that Win Shares, Games Started, Points per Game, and turnovers had the strongest correlations with salary. It was also interesting to note that Rebounds, Points, and Assists are all intercorrelated with one another, which can be explained by a change in the way basketball is played these days. Awhile ago, people could get by with being solely a scorer, or a pure rebounder or passer, but these days, in order to be a competitive player, you need to be good at all three of these things.

I was interested by the strong positive correlation between turnovers and salary. Wouldn't it be that the more you lose the ball, the worse of a player you are which leads to less pay? I did a quick scatter plot of turnovers per game and minutes per game and found that these two were also strongly correlated. My conclusion for this was that, if you are a good player, you will play and have the ball more, and if you have the ball more, there are more chances for you to lose the ball. So if a basketball player wants to earn more money, maybe he shouldn't purposely lose the ball, and instead work hard and play more minutes.



The last thing that I did during this stage of analysis was to see if there was a significant difference in average salary between the positions players played or between the teams they

played for. To do this, I made two sets of boxplots; one for the positions and one for different teams.



As one can tell from the plots, although there are many outliers in the point guard position, all five positions earn about the same median salary. As for different teams, I selected 10 teams to show on the boxplot, so a third of the total, and it is quite evident that some of the highest paid teams have much higher median salaries, but also higher salaries across the board. This may be caused by the NBA's soft salary cap, which enables teams and players to exceed the salary cap if they sign a multi-year deal or extension in order to promote "hometown loyalty". In the last section, I will check to see if these differences and correlations are statistically significant at a 95% confidence interval.

IV. The Statistical Exploration

I decided to do some tests to see if I can take these correlations and differences, or lack thereof, confidently. In order to test plausibility of the correlation coefficients between the different features. My hypothesis test for each of these read like this: The null hypothesis is that the pearson correlation coefficient for the population is equal to the one I obtained from the data, and the alternative hypothesis is that it is not equal to the one I obtained from the data. In order to check this, I shuffled up the data, keeping the pairs together, and split the pairs into 2

sets; one with 90% of the data, and one with the other 10%. From there, I split the pairs in the 10% of the data, and shuffled one of the features before repairing them back up. This would make sure to un-correlate that portion of the data. I then concatenate the shuffled 10% back on to the 90% and took the pearson correlation coefficient once more to compare it to the original one. I do this 10,000 times to see the likelihood of achieving a correlation as strong as the one from the data. I yielded p-values of 0.004, 0.015, 0.009, and 0.006 for points per game, win shares, games started, and minutes per game with salary, meaning for all 4 of these features, the empirical value of the correlation coefficient was accurate at a 95% confidence level. Because the distributions of these features were not normal, I checked my answers using the Mann-Whitney U test, a common test for non-normal distributions. I achieved incredibly low p-values from these tests on the order of -125.

The next thing that I did was check to see if the differences between the means of each team could be taken as truth, and not just with a grain of salt. My null hypothesis was that there would be no difference, while my alternative hypothesis was that there was a difference. I started with the two highest paid teams on average; the Golden State Warriors and the Oklahoma City Thunder. I found the difference in their means and did a bootstrapping test to see how often this difference in means might come up. To do this I took the salaries from both teams, put them into one array , shuffled them, and redistributed the salaries to the teams. If there truly was no difference between the two teams' salaries, then it would not matter that they were coming from the same set. For this particular test, I received a p-value of 0.466, which is higher than the alpha-level of 0.05 that I had set as my threshold, meaning I could not conclude that there was a difference in mean salary between the two highest paid teams. However, running the same test with the Warriors and the Kings, the highest and lowest paid team, I yielded a p-value of 0.026, meaning that I **can** conclude that there is a difference in pay

between the highest paid team and the lowest paid team, confirming what I had seen in my initial visual analysis.

The final thing I did was principal component analysis of the dataset, and found that when all things taken into account, ***there is only 1 principal component.*** I did not find this too surprising, since I saw during my EDA that many of the features had strong correlations with one another. Armed with this knowledge, I set out to perform an Ordinary Least Squares Fit to obtain some p-values. I took two different OLSes; one with the minimum amount of features, and one by removing the feature with the highest p-value until they were all reasonably low. For my first OLS, I looked at the 4 stat categories that I deemed most influential on the salary from my EDA and found that points per game and games played were the only two that weren't strongly correlated with one another. From this, I got a model with an r-square of 0.434 and both variables having low p-values. The other OLS gave me a regression model with games played, assists per game, points per game, win shares, and assists percentage. This second model gave me an r-square value of 0.504, so marginally better than the first one, but not as high of an F-statistic.

V. Concluding Thoughts and Revelations So Far

There are a few main things to take away from these initial findings so far. The first is that because of the way basketball is played these days, many of the features in this dataset are correlated with one another, leading to only 1 principal component. Some other things is that the correlations between points, rebounds, win shares, and minutes per game with salary are the numbers that I found in my initial EDA at a 95% confidence interval, and that at a certain level, there is a difference between the average salaries between teams, but not between any two positions.