Capstone Project 1: Exploratory Data Analysis

When I first began this project, I was initially focused on whether or not the Sacramento Kings would make the playoffs. I pulled team statistics from the past 5 seasons and individual player statistics from the past 10 seasons and began some initial exploration.

From the initial EDA of the team stats, there was a clear pattern amongst many of the variables indicating which teams made the playoffs from those that didn't. However, I was particularly fascinated by the individual player stats.

This was for a few reasons. Firstly, while basketball is a team sport, the talent level of each individual player directly contributes to that particular team's ability to win. Basically, the more talented your players, the higher the likelihood you are going to win games.

Secondly, looking at this from a team perspective is similar to looking at the economy from a macro perspective. You are able to assess a general direction but aren't really able to capture the small nuances that cause the economy to go that particular direction. The more I explored, the more I became interested in trying to assess an individual player's ability to contribute.

My initial hypothesis was that by examining each part of a player's game, we could highlight the skills that contribute to arguably the most important aspect of professional basketball: winning. Then I started asking myself questions like 'does a lockdown defender have a bigger impact on winning than a ball-handler?'. Or 'does being a better offensive player play a more significant impact than a defensive-oriented player?'.

Lastly, I had a lot more data to work with. For the team data, I could gather hundreds of observations while for individual players I could gather thousands, depending on how many seasons back I wanted to go. It seemed like a no-brainer and is why I decided to go the individual-player route.

When it comes to my exploratory data analysis, my focus was primarily on visualizing the forty or so potential explanatory variables and their relationship with our target variable, win shares per 48 minutes. Why did I choose this particular metric? Well for one thing 48 minutes is equal to 1 NBA game. We could've gone down the more macro route of trying to predict overall win shares (for the whole season) but I didn't think this would provide the same value as what a player did on a game-by-game basis. Additionally, I planned on comparing this number — i.e. win shares per 48 minutes — with the salary to generate a unique perspective on the value a player provided to a team compared to their cost (i.e. their salary).

One of the first interesting aspects of the data was in regards to the number of games played in. It was significantly left-skewed, with a sharp uptick around the 60 game mark. This potentially signaled that while teams have between 13-15 players on their active roster on any given night, that only a portion of these players actually get playing time.

When we took a look at the number of games started, there was further evidence supporting the previously stated 'portion of players' hypothesis. The overwhelming majority of players started exactly zero games! From there, the histogram quickly descends with a relatively small number of players starting between 20 and 70 games. At 82 games, there is a spike which gives us some evidence that teams could be using a core group of starters, leading to only a portion of players playing in any given game.

Now when creating the first scatter plots — focusing on win shares per 48 minutes and minutes played, field goals, and 3-point field goals — there was a significant distortion of the graphs due to outliers. To address this, we used the PER statistic which stands for player efficiency rating.

Now why use PER for this? Firstly, because it is somewhat similar to win shares in that it uses both offensive and defensive statistic to give a holistic picture of a player's overall skill set. Secondly, it had a statistical cut-off in that players had to have played 6.09 MPG (minutes per game). After instituting the PER cut-off the distortion was for the most part eliminated. I want to note however that I didn't institute the second part of the cut-off which was that the player had to have played in more than 500 minutes total that season as well. The primary reason for this was due to the fact that it would have eliminated approximately 46% of my data set if I had and I did not want to eliminate so much data for what at best would be a marginal return.

That being said there were some interesting takeaways from the exploratory data analysis with some being expected and others being rather unexpected/disappointing.

The first, and perhaps the most obvious was the correlation between shooting and win share. Field goals made, field goal percentage and true shooting percentage — which essentially combines field goal percentage, free throw percentage, and three-point percentage into one — all appeared to have a strong positive impact on win share. Basically, the more you shoot and the more shots you make, the higher your likely contribution is going to be towards your team winning.

The second takeaway was that assists and rebounds don't appear to have any significant relationship with a player's win share. This was rather surprising since outside of scoring these are perhaps the two most common statistics used to describe how 'good' a player is. Additionally, it is generally assumed that better teamwork generally equals more wins but we're not seeing that to be the case at least with these particular graphs. With regards to rebounding, if you squint just right there may appear to be a slight positive correlation with win share but overall I'd say their relationship is ambiguous at best.

The third takeaway was that offense is king. Most of the variables associated with defense — like steals and blocks — returned as indiscernible blobs with no relationship whatsoever with win share. After seeing the relationships with offensive variables and shooting in particular, being a good offensive player who can shoot the ball well looks to have more of an impact on winning than being a good rebounder who can block shots.

In a way this makes sense. A hypothesis I have is that while grabbing steals or blocking shots may lead to a team gaining additional possessions, the rarity of these respective statistics (averages usually hover around two per 100 possessions for both) cancels out any positive impact they may have.

The fourth and perhaps most important takeaway is the potential for multicollinearity between variables. While I didn't technically test for this, there are quite a few variables that interact with each other. A prime example of this would be the three variables associated with FG — field goals made, field goals attempted and field goal percentage. The three have a very intricate relationship, with making more field goals (or shooting more) having a direct impact on field goal percentage.

There are a few that may not be as obvious, especially to those unfamiliar with basketball statistics. One is PER — player efficiency rating — which is calculated by combining numerous offensive and defensive variables together. There are weights given to each but we'll have to be careful when it comes time to determine which features to use for our machine learning algorithms.