My biggest epiphany at the end of 2022 was the power of taking frameworks and teachings from one industry and using them to have a different perspective in another industry. I first noticed this while studying the backgrounds of the world’s best investors for my day job as an institutional investor. From Robert Smith with engineering to Warren Buffett with horse race betting, the most consistent way I’ve seen to create an “investment edge” is to apply a way of thinking from another domain to investing.

Given my love for basketball, I started to work on asking investment-related questions about my basketball game. The questions include “How can I increase the Sharpe Ratio of my game?” and “How can I ensure there isn’t too much positive correlation amongst my teammates when choosing players in a pickup game setting?”.

For this project, I will ask about similar stock market concepts when evaluating how well a team is constructed after reviewing past playoff advanced statistics results. One of the first lessons I’ve learned as an institutional investor is that outperformance tends to come during down markets, so you should put more weight on how well a manager does in bear markets or when their investment style is out of favor. I’ve taken that view to look at the NBA (and almost all sports) in that the style of play in the playoffs matters exponentially more than during the regular season. Most of the basketball analytics community doesn’t have this belief, but I think part of that view stems from having more data to work with for a regular season (82 games for 30 teams) than for the post-season (6-12 games for 16 teams). However, if the goal is to win a championship, a general manager should construct a team to succeed in the playoffs.

For the first part of the project, I want to find the “market inefficiency” between what features contribute to winning in the playoffs and overall salaries. For the second part, I will create the factors needed to develop a Discounted Cash Flow (DCF) model for estimating a player’s “intrinsic value” using the same playoff statistics.

In this project, I will also introduce one of the key concepts I learned while pursuing a Master’s in Economics: sub-classification is power for prediction purposes. I’ll primarily focus on using OLS regressions for prediction since it is the most straightforward model to interpret, although I will introduce more complexity along the way.

Now, let’s get started with the project! I’ll start by loading all the packages I need for this project.

**Data Collection and Wrangling**

I wanted to use three data sources for each season: advanced playoff statistics by player, annual salary by player, and salary cap. I will start by collecting the advanced playoff statistics from basketball-reference.com. With some assistance from ChatGPT, I will be scraping annual advanced statistics by player. These statistics include a player’s True Shooting Percentage, Usage Rate, Wins Share, and Box Plus/Minus. Although I originally wanted to cover over 20 seasons with this project, I had to limit the scope due to the availability of salary data (more on this later).

I had to get creative to pull the salary data for each player. The best resource I found for this was espn.com (the link to the webpage for the 2020 season is below), but I couldn’t figure out how to scrape the data from a table with multiple “pages within a page” for one season.

After some serious Googling, I came across Unboxed Analytics, where the author of the blog (Erik Webb) scraped the data I needed with R (https://unboxed-analytics.com/data-technology/web-scraping-nba-salaries/). Given that I couldn’t find the equivalent Python packages for the R packages he used (rvest, tidyverse, and stringr), I just downloaded the result and used it for this project. Thanks, Erik!

This dataset became the “bottleneck” for me, as the data from ESPN for the NBA became sparse before the 2006 season. Additionally, the Unboxed Analytics dataset only went through the 2020 season. These were the new data constraints that I was operating with.

[www.espn.com/nba/salaries/\_/year/2020/seasontype/3](http://www.espn.com/nba/salaries/_/year/2020/seasontype/3)

Given the NaN for Darrell Armstrong in the dataframe, I assumed that several players had data missing, especially the further back in time I went in time. To deal with this, I’ll remove observations that are missing values from specific columns. Since the columns from the player salary data have the lowest number observations, I’ll start there and repeat until I get to a point where the number of observations is the same across all features.

After two iterations, I have 2,586 observations for each feature. I also realize that I need to change the data type for most of the features and remove a few duplicate columns.

Now that I have cleaned and merged the data, it’s time to explore the data further and add some more features that I think are important for this project.

**EDA and Feature Engineering**

I’ll start by understanding Win Share per game (WS/48), as this feature will be how I judge how a player contributes to winning.

Wins Share per game visually looks normally distributed, which is ideal given that I’ll utilize linear regression later. Let’s confirm this view with descriptive statistics.

The descriptive statistics show that WS/48 is close to being normally distributed, although there are still higher moments (positive skew and extremely positive kurtosis) to consider. I’m also curious about how this feature looks for different positions on the court. Time to view more histograms!

The WS/48 for each position tends to fall within the 0.0-0.3 range, with a few significant outliers to the positive. The position this doesn’t hold for is shooting guards (SG), which I’ve always viewed as the position least likely to contribute to winning based on the role of most shooting guards (i.e., primary focus on scoring alone) within a team structure. The high kurtosis is coming from outliers happening more than expected, which I will correct for later.

One area I want to dive deeper into is how minutes per game (MPG) influence wins per 48. I will create a categorical variable for players who play “starter” minutes, are “in the rotation,” and barely get on the court. But first, let’s create a feature to track MPG in the first place.

*Will decide what to do with front/backcourt*

Given that non-GT (or garbage time) players disproportionally make up the number of observations relative to their production, I will revisit the distribution of wins per 48 without the GT players.

Although the distribution for the non-GT players is closer to normal, there are still higher moments (mainly positive kurtosis) that I would like to further account for. I’ll see if winsorization or setting absolute values (maximum or minimum) would be a better approach to handling the higher moments.

Winsorizing led to a better distribution, so I’ll use this method. I’ll also see how WS/48 looks by position to see where the outliers (if any) exist.

These results look promising! Let’s also view the same feature while parsing out the rotation status of the various players.

Looking at the strip plot above, I’m most surprised by the distribution in garbage time (GT) WS/48. I would expect more of a cluster near zero, as these players don’t have enough time on the court to make a significant impact.

One of the beliefs I have for championship-caliber teams is that they should be able to win about two-thirds of their games in the regular season and the playoffs. This standard would lead to at least ~53 wins in the regular season and winning a playoff series in six games (4-2). Winning that many games in the regular season is typically enough to get good playoff seeding while not over-extending star players during an 82-game season. Based on this belief, I’ll reshape the WS/48 target feature to reflect the win percentage for a 65% win rate. I will also minimize the impact of GT player’s contribution to winning.

Although there are more outliers in the data, I attribute that to minimizing the impact of GT’s win production. Seeing the most substantial negative outliers at the Center position is also surprising. I hope those players never played for my Lakers!

Another change I want to make is to the salary cap. First, I want to expand the salary cap, as most playoff teams tend to have payrolls that are 20-40% higher than the salary cap for any given year. I also want to compare a player’s salary to the expanded salary cap. This desire will lead to a new feature (cap\_percent) to compare to a player’s percentage of the desired winning percentage (WS/48\_percent). If a player’s production percentage is higher than the salary cap percentage, then a team would have found “value” in that player.

It looks like Centers and Power Forwards cost the most to acquire when based on the median value. I attribute part of this to fewer people in the world being tall enough (usually 6 foot 9 inches and above) to play those positions. This result also leads to one of the aspects I want to test: the influence of positional value later on.

Another area where I think taller players can provide more value (or “perceived value”) is on the defensive end. This view also supports another hypothesis I have for this project: championship-caliber teams tend to have players who are not liabilities on either end (offensive or defensive) of the court.

I’ll start exploring this hypothesis by creating features related to contributing or detracting on offense or defense. I’ll start by calculating how much of a player’s value comes from the defense using the defensive plus/minus statistic.

My initial view that Centers and Power Forwards have more impact on the defensive end is holding up, especially for starting players. The distribution is a little wonky though, as some player’s value may come entirely from the defensive end. I’ll adjust the distribution later on in the project.

Using the plus/minus statistics, I’ll create categorical features to represent whether a player is an asset or a liability on both sides of the court. Since I’m more focused on liability, I’ll create a feature for this first.

Surprisingly, more players were offensive liabilities than defensive liabilities. My initial thought is that rotational status plays a role here. I’ll test that view below.

My initial inclination was correct, as the difference between the starters and bench players is that bench players are likelier not to have the consistent offensive skills to stay on the court. It also speaks to bench players needing a different skill set than starters on a team.

I’ll now focus on creating categorical features for “asset” players on the court. What’s more interesting to me are the players that are “assets” on both ends of the floor and star players (based on a higher usage rate and rotational status) that I would consider an asset. Just like liabilities can have a non-linear negative influence on wins, I believe “star assets” have a non-linear positive influence on winning.

Overall, there aren’t many “star assets” in the NBA. Based on the figure above, I calculate an average of 7-8 star asset players a season.

Now that we have the asset and liability categorical features out of the way, I will create the last categorical feature based on a player’s age. There’s been a lot of work on when a player’s production peaks (around 24-25 years old), while the NBA’s Collective Bargaining Agreement (CBA) increases the veteran’s minimum the longer a player is in the NBA. Before creating the feature, I will see where I should place boundaries regarding a player’s increasing and decreasing production value.

Based on the information in the graph above, I’ll mark a player’s ‘prime’ for wins production between the ages of 24 and 30.

I’ll now look at the distribution of the other numerical features used in this project to understand the underlying characteristics better.

There are a lot of positive outliers in these features, plus the defensive add feature still looks weird. Given the heavy reliance on linear regression, I’ll winsorize the data for 90% of observed values utilized. The winsorization will allow for some outliers to persist without them having too much of an influence on the regression results. I’ll also reset the minimum value for the features to a percentage near zero (0.0001). The code for this section is another area where ChatGPT was extremely helpful.

The distribution of these features is better suited for linear regression analysis. I will also add more features related to a “risk/reward” view for player statistics. An example of this type of statistic is the assist-to-turnover ratio.

I now have all of the features I wanted to use for this project. I’ll review all of the features and remove the ones that will not be useful from this point forward.

**Part 1 – Finding Market Inefficiencies between Wins Production and Salaries**

This section aims to understand better which independent features have a linear relationship to the target features (win percent and salary cap percent) and see where there are disconnects between the two target features. Before diving deeper, I’ll list a few hypotheses I have about both target features and how they’re related:

1. Based on playing thousands of pickup games and watching the NBA playoffs, I believe that the team that has the fewest weaknesses (combined with a go-to scoring threat) tends to be the best team. Based on this view, the offensive and defensive liability categorical features would be statistically significant features.

2. Although true shooting percentage is important, it’s more important based on the makeup of shots a player takes. This is especially important regarding a player’s free throw rate, as free throws are both a positive for the offense and a foul for the defense. Austin Reaves’s game for the LA Lakers is a perfect example of this, as his ability to get to the free throw line plays a significant role in him being a positive contributor to winning while being the third option on a playoff team. The free throw rate to true shooting percentage ratio (FT\_to\_TS) would be statistically significant and positive if correct.

3. Although assists are essential at the team level, I don’t think they’re that important on an individual player level. What’s more important is the turnover rate for a player relative to their usage rate. This is why I think players like Trae Young, Lemalo Ball, and the New York Knicks version of Jeremy Lin were all overrated, as they had high usage rates to go along with their turnovers. If correct, the assist rate would be statistically insignificant, while the turnover rate to usage rate ratio (TOV\_per\_USG) would be statistically significant.

4. From Lebron James, to Larry Bird, to Jayson Tatum, to Jimmy Butler, I view Small Forward as the most important position on the court. This is the only position that can guard all five positions and bring the ball up the floor to initiate offense from a conventional sense. Based on the EDA, this also gives me the view that the Shooting Guard position may also be overrated. If correct, a player would be a statistically significant contributor as a Small Forward while being a statistically significant detractor as a Shooting Guard.

I’ll start this section with a correlation matrix heatmap while utilizing the assistance of ChatGPT.

The most significant observations I make from the correlation matrix heatmap are 1) true shooting percentage (TS%) has the most immense impact on win percentage out of the advanced metrics I’m comparing, 2) minutes per game (MPG) plays a significant role in both win percent and salary cap percent, and 3) there is almost zero correlation between win percent and salary cap percent. I will now conduct a variance inflation factor (VIF) analysis to see if multicollinearity exists in the data, focusing on different rebounding, assist, and shooting efficiency measures.

The ideal situation is to have the VIFs of the independent features in question be less than 5. Unfortunately, all three rebounding measures are above this figure. However, I do believe that being good at rebounding on one end of the floor doesn’t fully translate to the other end. The VIFs also point this out, with total rebounding (TRB) having the highest score. I’ll see how the scores change if I remove the TRB feature.

The VIF scores are much lower after removing the TRB feature, so I’ll exclude it for the rest of the statistical analysis. I will run a series of linear regressions for both win and salary cap percentages as the target features. I’ll start with one regression for each that considers the advanced metrics of interest with the categorical features.

Based on the first regression, the true shooting percentage (TS) is the most significant feature for determining a player’s win contribution based on the t-statistic. Based on the initial correlation matrix, I saw this coming, but it’s also nice that most of my hypotheses were correct. Both liability features were negative and statistically significant, the free throw rate relative to true shooting percentage matters more than the free throw rate alone, the assist rate is statistically insignificant while the turnover per usage rate is significant, and the Shooting Guard position is a negative contributor to winning. The only hypothesis I was incorrect on was the value of the Small Forward position.

Other prominent features include offensive asset, defensive liability, defensive asset, minutes per game (MPG), and offensive rebounding percentage (ORB). Other interesting observations from this regression are 1) a player being an offensive asset matters more than the liability categories, 2) block and defensive rebounding percentages aren’t that important while offensive rebounding is important, and 3) the intercept starts off with a negative number.

I’ll admit to being annoyed that there are still multicollinearity issues in the data, even though I tried to minimize this from happening. I have one more trick up my sleeve to fix it, but I’ll examine the regression for the same features against the salary cap percentage first. Here are a few initial hypotheses:

1. Given that sports economists have already proven that teams overpay for scorers while underpaying for non-scorers, I expect the same to hold here. If correct, this insight means offensive categories (usage rate, offensive asset, and star asset) will be statistically significant contributors. In contrast, top-down defensive metrics (defensive-add and defensive asset) will be statistically significant detractors.

2. Following up on this view, backcourt players’ (Shooting Guards and Point Guards) usage rates tend to be highest. If correct, then both positions would be positively statistically significant.

3. A player tends to be evaluated based on the three main box score stats that can make up a “triple-double”: points, rebounds, and assists. Since it’s possible to get more defensive rebounds than offensive, the defensive rebounding and assist rates would be statistically significant, while the offensive rebounding rate would be insignificant if correct.

4. Given the Collective Bargaining Agreement (CBA) structure, older players will always get paid more. Based on this view, age and related features (prime status) will be positively statistically significant.

For this regression, minutes per game is the most significant feature based on the t-statistic, although this is also expected. Once again, most of my hypotheses were correct. The incorrect ones were the backcourt argument and offensive assets being positive. There are also some interesting aspects to point out when compared to the regression on win percentage:

1. True shooting percentage is a negative contributor to salary percentage but a positive contributor to winning. Vice versa is true for the usage rate.

2. Turnover-related statistics (TOV, AST:TOV, TOV\_per\_USG) are insignificant for salary percentage but are significant for winning percentage.

3. Older players cost more, although it is uncertain how much they contribute to winning.

4. The AIC/BIC were lower for the salary cap percentage regression than for the winning percentage regression. This result means that the salary cap percentage regression is better for making future predictions, even though the winning percentage regression captured more in-sample variation with a higher R-squared. I contribute this to the salary cap percentage regression having fewer insignificant factors while believing that having more statistically significant categorical features leads to a lower AIC/BIC.

I’m still annoyed by the multicollinearity issue, so I’ll standardize all continuous independent features to see if this issue goes away. I’ll do this by turning these features into z-scores. Although it makes interpreting more challenging, I can still understand the key features. I’ll do this step and then re-run the first regression on salary cap percentage to see if any results change.

Moving to z-scores removes the multicollinearity issue while changing the intercept from negative to positive for the win percentage regression. I’ll now see if the same result happens for the salary cap percentage regression.

Moving to z-scores removes the multicollinearity issue while changing the intercept from negative to positive for the salary cap percentage regression as well.

One of the econometric topics I wanted to test with this project is how sub-classification may alter the results of which independent features are significant. Although this introduces the potential “curse of dimensionality,” I believe sub-categories should be treated differently and could lead to discovering more inefficiencies. An example of this in the investment world is determining which factors drive company financial results in different sectors. Inventory turnover matters more for consumer-based companies, customer retention matters more for software companies, and the oil price matters more for energy companies. I want to explore two sub-classifications: rotation status (starters and bench players) and prime status.

I’ll start with rotation status, separating the data into the different sub-classes, then running the regressions for the target features to see the differences from utilizing all observations.

Some initial observations with the win percentage regression for starters only vs. for all players include 1) a higher intercept value, 2) the R-squared is higher, but the AIC/BIC is lower, 3) no position is statistically significant, 4) the usage rate and risk-adjusted features are all statistically insignificant, and 5) several metrics that weren’t statistically significant before (STL, BLK, DRB, AST) are now significant. The most prominent feature that changed is assist percentage (AST), which is now the second most important feature based on the t-statistic. So much for my “assists are overrated” hypothesis.

I’ll now run a regression for the salary cap percentage for starters.

Some initial observations with the salary cap percentage regression for starters only vs. for all players include 1) a lower R-squared and AIC/BIC than the salary cap regression on all players and the win percentage regression for starters (which is the expected outcome), 2) a much higher intercept value, 3) Power Forwards becoming an insignificant position, 3) true shooting percent metrics (TS and Three\_Pt\_to\_TS) also becoming insignificant, and 4) free throw rate (FTr) becoming positively statistically significant. Outside of the lower R-squared and AIC/BIC, the feature comparison between the win percentage and salary cap percentage regressions for all players are upheld for starters only.

It is time to run more regressions, focusing on the bench players.

Positions finally matter; every position listed has positive statistical significance, and backcourt players matter the most. This insight means Centers not good enough to start for a team are more likely to not contribute to winning. The true shooting percentage also matters much more, as the coefficient is 2.5x higher than with the win percentage regression on all observations. My initial hypotheses are breaking down more, with offensive liability and most risk-adjusted features being statistically insignificant while the assist rate is significant. Lastly, minutes per game have now become insignificant. I believe this is because most bench players have similar minutes but produce wins at very different rates.

I’ll now run the salary cap regression for the bench players.

Based on the R-squared, the regression fitting the salary cap percentage for bench players has been the worst to this point. My age-based hypothesis is also the only theory that holds up for paying bench players, as the traditional counting stats aren’t as helpful, and a bench player is less likely to be a major offensive contributor. One surprise is the positive influence of the Three point rate, leading to statistical significance. Then again, the Three point rate adjusted to the overall true shooting percentage is negatively statistically significant at a similar coefficient, which could lead to the two features canceling each other out.

My last set of regressions is based on sub-classifying the data based on prime-age status. Since I don’t see the value in viewing the target features for garbage time (GT) players, these regressions will better capture all of the observations while helping me understand the value a player can bring to a team as they age. I’ll start with the players in their prime (age 24-30) and fit a regression for their win percentage.

**Part 2 – Creating DCF Inputs**

**Part 3 – Finding “Alpha” in Wins Production**

**A screenshot of a computer

Description automatically generated**