**NBA Roster Construction with STAR: SRS Team Adjusted Rating**

Track: Basketball

ID: 7V7B0BZ38

**Abstract**

The use of analytics for NBA Roster construction has become more prevalent in the last decade. Front offices now look at advanced player statistics like Box Plus/Minus (BPM) and Player Efficiency Rating (PER) when making trade and free agency decisions. While these advanced metrics help teams make effective roster decisions, quantitatively projecting the degree to which a specific move affects a team is difficult. In other words, it is usually easy to answer the question, "Does this move improve my team?" but it is tough to answer the question, "How much does this move improve my team?". Our model, STAR, aims to answer the latter question.

STAR, or SRS Team Adjusted Rating, predicts a team's SRS (Simple Rating System) based on its players' roles and statistical summaries. In order to summarize player skillsets, we made use of PCA (Principal Component Analysis) compression on relevant per-game and advanced statistics (gathered from Basketball-Reference.com) to create 5 Factor summary statistics for each player. Instead of using traditional positions to define player roles, K-Means clustering of the same data was used to group players into seven archetypes. These archetypes and the 5 Factor summaries provide a simplified view of a player's skill set while at the same time retaining enough granular information to remain accurate and useful. Archetypes and 5 Factor summaries for every NBA player dating back to 1985 were then combined to form vectors of player information for each team year. Finally, these vectors were run through a random-forest decision tree model to predict team SRS. The STAR model provides analysts with multiple methods of analyzing ideal roster construction and gives teams the ability to build a roster from scratch and accurately predict their SRS.

1. **Introduction**

Coming into the 2019-2020 season, Las Vegas oddsmakers set the Philadelphia 76ers over/under win total at 54.5, the second highest in the league (Basketball-Reference). The 76ers were +775 to win the championship, the fifth-highest odds in the NBA (Basketball-Reference). When these projections came out, no one batted an eye, as Philadelphia was a team on the rise. The previous year they had fallen in the Eastern Conference Semifinals to the eventual champion Toronto Raptors, losing on one of the most iconic Game 7 shots in NBA history. The team brought back their two young superstars in Joel Embiid and Ben Simmons and a fringe All-Star candidate in Tobias Harris. In free agency, Philadelphia brought in Al Horford, a former All-Star and respected defensive veteran. Even though they lost Jimmy Butler to the Miami Heat, they brought back Josh Richardson, a solid 3 and D player, as part of a sign and trade deal. On paper, this team looked like a defensive powerhouse with offensive firepower.

However, as the season progressed, it became clear that this Philadelphia team would not live up to its billing. Their defense was solid, but a 109 defensive rating was not at the elite level expected entering the season. The offense was only middle of the pack, and it was clear that the team was struggling with spacing issues. The 76ers ended the year with an SRS of 2.25, which equates to 47.6 expected wins, a significant drop from their preseason projected win total. The 76ers disappointing performance begged the question, what happened?

Looking at Philadelphia's roster construction, a couple of things became clear. Subpar 3-point shooting seasons from Harris, Richardson, and Horford (all under 37% from 3) magnified the offseason loss of JJ Redick (45% from 3) and further compounded Simmons' and Embiid's 3-point struggles. The team also lacked secondary playmaking behind Simmons, who often struggled when forced to play in the half-court. Generally speaking, while it seemed like Philadelphia added talent in the offseason, it appears the questionable fit of their additions with their franchise cornerstones led the team to underperform. However, recognizing the poor construction of the Philadelphia 76ers roster is not a difficult task. Any fan with enough experience watching the team could tell you the pieces did not work together. The real question then becomes: how does one build a roster the right way?

The question of ideal roster construction strategy has long plagued NBA front offices. Should executives prioritize stars or depth? Talent or fit? How can front offices determine the right players to pursue? We noticed that there were very few methods for a front office to quantitatively analyze the roster decisions that they were making, especially methods that took player fit into account. This paper introduces STAR, SRS Team Adjusted Rating, a model that predicts a team's SRS based on its players' roles and statistics. SRS, or Simple Rating System, is a team's average point differential adjusted for strength of schedule. We believe that it is the best single number estimator of a team's quality. In the last 74 years, 50% of the NBA champions led the league in SRS that season, and the league champion is rarely outside of the top five. By predicting SRS, our model gives front offices a concrete idea of how well their team will perform over the course of the season.

STAR allows front offices to build a team from scratch and evaluate the quality of the inputted group of players. The model uses unsupervised machine learning to cluster each player into archetypes that describe a more detailed role than traditional positions can. These archetypes are then linked with compressed 5 Factor summaries of the player's statistics to create a numerical profile for each player. Finally, these profiles are combined to form an information vector for the team, which is passed into a random forest decision tree. The tree is trained on similar information vectors for every team dating back to the 1985 season. Based on what the model has learned from historical combinations of archetypes and statistics, it will predict the inputted team's SRS.

We use the following questions as a structure for the paper:

* Can a player's skillset be summarized by certain factors and clustering? (Section 3)
* How accurately can we predict a team's success based on players' archetypes and statistical footprints? (Section 4)
* What are ideal team building principles? (Sections 5, 6)

1. **Data**

**2.1 Data Collection and Pre-Processing**

We manually scraped 35 years (1985-2019) of player data from multiple Basketball-Reference tables. The tables were merged together based on player name and season. Our dataset contains 18,734 observations of 51 variables.  The variables are a combination of rate statistics, advanced metrics, shooting efficiency statistics, and counting statistics. Each row corresponds to a player in a given season.

**Figure 1:** *Minutes Played vs. Box Plus Minus*

Chart

Description automatically generated

Figure 1 displays a scatterplot of a player's Minutes Played over the course of a season, plotted against their BPM. Because BPM can be unreliable in small samples, the figure shows that most of the BPM outliers are from players that played less than 100 total minutes. These players' performances in small samples lead to unrealistic advanced statistics that can bias the model. Due to this concern, we filtered our data to contain only players who played more than 100 minutes in a given season. With this filter, the dataset shrinks to 16,190 observations. The new data has less drastic outliers, which helps reduce one possible source of bias in the model. Figure 2 below shows the updated Minutes Played versus Box Plus/Minus scatterplot.

**Figure 2:** *Minutes per Game vs. Box Plus-Minus (At Least 100 Minutes Played)*

Chart, scatter chart

Description automatically generated

**2.2 Variable Selection**

To effectively summarize players through clustering and PCA analysis, it was imperative to select a range of statistics that accurately describe a player's contribution. To accomplish this task, we selected a combination of rate statistics, advanced metrics, and counting statistics. We chose 38 statistics that attempt to holistically measure a player's impact as well as their play style. The table below defines these 38 variables.

**Table 1:** *Description of Variables Used in Analysis*

Table

Description automatically generated

**3. Clustering and Five Factors**

**3.1 Cluster Creation**

In order to judge the composition of an NBA roster effectively, we first needed to judge and summarize a player's abilities. However, using 38 statistics for every player on every roster in the NBA would create overly complex and incomprehensible data. As such, we made use of PCA (Principal Component Analysis) reduction and K-Means clustering in order to provide brief, accurate summary reports detailing a player's on-court contributions.

K-Means clustering is an unsupervised machine learning method whose goal is to classify data points into K unique clusters. Once the value of K has been selected manually, the k-means method continually updates the cluster centroids and resulting cluster assignments in order to minimize the within-cluster variances. Each point is assigned a cluster label according to the cluster centroid to which it has the shortest Euclidean distance. Often, K-Means clustering is combined with PCA reduction, which reduces the dimensionality of the data while still retaining most of the distinguishing information. In our case, the 38 statistics we chose as variables were reduced to a dimensionality of 15 components. Although we lost 23 dimensions in our new dataset, the new data was still able to explain 95.1% of the variance in the original data. These 15 principal components were then passed to the K-Means algorithm for cluster assignment.

The next step was deciding the optimal value of K to ensure accurate and useful clustering. In many data sets, a simple visual representation is enough to identify a clear number of distinct clusters, and thus the optimal value of K. NBA player statistics, however, proved to be quite diverse. Only 55.7% of the variance can be modeled within 2 dimensions and 64.1% within 3 dimensions. The visual representations of the PCA condensed data in these dimensions show no signs of a clear K value (Figures 3 and 4).

**Figure 3:** *2-Dimension Visualization of NBA Player Statistics*

Chart, scatter chart

Description automatically generated

**Figure 4:** *3-Dimension Visualization of NBA Player Statistics*

Chart, scatter chart

Description automatically generated

To select the number of clusters, we utilized inertia score analysis, another common method of K selection. In essence, inertia scores range from 0 to infinity and measure the distance between points within a cluster. Lower scores represent tighter and more useful clusters. Inertia scores can be used to select the optimal K value by performing PCA reduction at each of the possible dimensionalities and graphing the inertia scores at each dimensionality level. The optimal point is chosen at the elbow point, a term used to represent the point at which the inertia scores begin to slow in their descent, and the graph becomes nearly linear rather than polynomial/exponential. This point is considered the optimal number of clusters because it balances model complexity with diminishing returns in cluster inertia. Within our data reduction analysis, the elbow point appeared around a dimensionality of 7, indicating an optimal clustering with 7 unique groups of players (Figure 5). This clustering is shown (in two dimensions for visibility) in Figure 6.

**Figure 5:** *Inertia Plot*

Chart, line chart

Description automatically generated

**Figure 6:** *K-Means Clustering (K=7) on 2-Dimensional Data*

Chart, scatter chart

Description automatically generated

**3.2 Creating the 5 Factors**

In past papers on the topic, researchers decided that some form of clustering (K-Means or Gaussian Mixture Model-based Clustering) was enough to provide accurate representations of a player's skillset (Kalman, Bosch). However, we believed that just a single label (or probability distribution between cluster labels) did not provide an accurate description of a player's statistical footprint. Of course, the overriding principle in these original decisions was the idea that a model can only provide so much detail before becoming overly complex, resulting in difficulty of use and comprehension. However, we believed that responding with the oversimplicity of a single classification system was too extreme a measure taken in the opposite direction.

To balance model complexity with detail, we performed a second PCA reduction without clustering in addition to the cluster assignment. The goal of this second reduction was to provide a multi-component summary report which could explain large amounts of variance in the data without losing its simplistic nature. We revisited the explained variance analysis that was done in the clustering stage to determine the optimal dimensionality value for our reduced data set. In the end, we settled on a dimensionality of 5 for our newly condensed data. 5 dimensions explain 74.7% of the variance in the original data and is still small enough to be easily digestible when attempting to extract meaning from inherently nameless PCA component vectors.

**3.3 Interpreting the 5 Factors**

As noted earlier, while PCA reduction is incredibly useful for scaling down data to manageable sizes, it loses comprehensibility and meaning as each of the new principal components is a weighted combination of the original high-dimension variables with no inherent name. However, because we deemed user-understanding of the inner workings of the STAR model to be highly important, we analyzed each of the 5 Factors to understand what traits and/or combinations of traits the factor represented. While some factors showed very strong correlations with certain variables from the original data (Figure 7), others seemed to have no relationship at all (Figure 8).

**Figure 7:** *Factor A versus Points*

Chart, histogram

Description automatically generated

**Figure 8:** *Factor D versus PER*

Chart, scatter chart

Description automatically generated

To analyze the extent of each of these relationships, we made use of a correlation matrix between the new 5 Factors and all of the original numeric data (Figure 9). By identifying strongly correlated factors (both positive and negative), we were able to label each of the 5 Factors with relevant real-world skills (Table 2). However, it is important to note that the additional variance explained by each factor decreases as more components are added. As such, it is not surprising to see many stronger correlations in Factor A than Factor E.

**Figure 9:** *Correlation Matrix Between 5 Factors and Original Data*

Chart, bar chart

Description automatically generated

**Table 2:** *Correlation Analysis of 5 Factors*

Table

Description automatically generated

**3.4 Player Examples of 5 Factor Summaries**

Now that we have an idea of what each of the 5 Factors represents, we can look at a few examples of real NBA player summaries for better understanding.

In Figure 10, we see the 5 Factors summary for Dallas Mavericks Center Kristaps Porzingis. Clearly, Porzingis is a gifted offensive scoring weapon who seems to have a decent interior focus. Using our prior knowledge of Porzingis, we can confirm this assessment, given his history of high volume scoring, rebounding, and blocking shots. His Inefficiency seems to be around the middle of the pack, and his perimeter defense score is among the worst in the league. Once again, this matches up with our prior knowledge, as Porzingis has a habit of taking long contested two-point jump shots, which decreases the efficiency of his strong scoring game. Additionally, he is a 7'3" center with limited lateral agility, explaining his weak perimeter defense grade. Finally, Porzingis himself was the first player to be nicknamed "The Unicorn" due to his combination of size and shooting ability. Fittingly, he grades above the 99th percentile in the Unicorn factor.

**Figure 10:** *Kristaps Porzingis 5 Factor Summary*

A picture containing graphical user interface

Description automatically generated

A second example can be seen in Figure 11: the 5 Factors summary for Miami Heat Shooting Guard Jimmy Butler. Once again, Butler ranks highly in the offensive talent factor given his robust scoring game. His slashing guard style results in the middle of the pack interior focus, and his high Inefficiency grade likely results from his diet of inefficient midrange shots and reluctance to shoot three-point shots. Unlike Porzingis, though, Butler is well known for his defensive acumen, hounding perimeter ball handlers with relentless passion, as evidenced by his 85th percentile perimeter defense grade. His unicorn grade is in only the 8th percentile, a sign of his old-school throwback style.

**Figure 11:** *Jimmy Butler 5 Factor Summary*

Chart

Description automatically generated

**3.5 Interpreting the 7 Archetypes (Clusters)**

After adding descriptive labels to each of the 5 Factors, we decided to do the same for the 7 player archetypes identified by the K-Means clustering. To perform this exploratory data analysis, we compared the distributions of each of the 5 Factors within each of the 7 archetypes. In order to give the values more meaning, each of the 5 Factors was scaled to have a mean of 0 and a standard deviation of 1, creating z-scores. Additionally, we viewed the raw clustering data, sorted by distance from the cluster centroid, to see examples of NBA players that fit each archetype best. For example, in archetype A, the distribution of the 5 Factors can be seen in Figure 12.

**Figure 12:** *Distribution of 5 Factors within Archetype A*

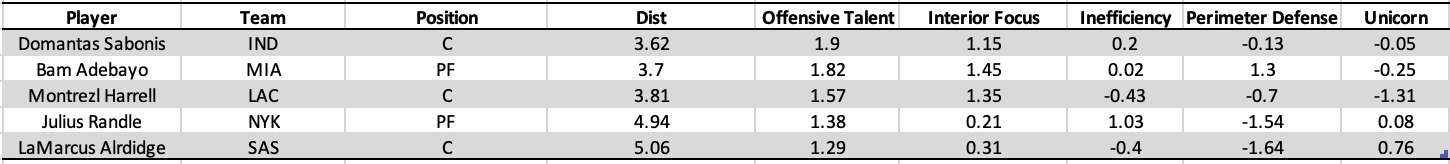
Chart, box and whisker chart

Description automatically generated

These results indicate that players in archetype A are very good offensive scorers, with high interior focus, medium to high Inefficiency, and poor perimeter defense. The unicorn factor seems to have a wide range of outcomes, which makes sense given the uniqueness of its skill combinations.

By looking at the players in this archetype (viewing only 2019 players for ease of comparison), sorted by their distance from the cluster centroid, we can confirm the previous findings given our existing knowledge of the players and their skillsets. The 5 most representative players for archetype A in 2019 are displayed in Table 3.

**Table 3:** *Players Representative of Archetype A with their 5 Factor Scores*



Given the distributions of the 5 Factors for the entire archetype and the most representative players according to the distance metric, we decided to name the archetype "Skilled Big." This process was then repeated for the other 6 archetypes (Figure 13 and Table 4).

**Figure 13:** *Distribution of 5 Factors within all Factors*

Chart, box and whisker chart

Description automatically generated

**Table 4:** *Archetype Descriptions*

(Green indicates positive correlation, Red indicates negative correlation)

Table

Description automatically generated

**4.**   **Random Forest Implementation**

**4.1.   Introduction**

After performing cluster analysis and creating 5 Factor summaries for every player in the league, we wanted to see how combinations of players would impact team success as a whole. The next step in the process was to create the STAR model: a mapping from player summary statistics to team SRS.

**4.2.   Data**

In order to summarize a player's statistical footprint, we made use of 11 factors. 6 of the factors come from our previous analysis, the archetype, and 5 Factor summaries. An additional factor comes in the form of organizational data: the player's name. The final 4 variables included in the final model are VORP, OBPM, DBPM, and MP. While each of these last four statistics are baked into the archetype and 5 Factor summaries, we felt it was especially important to single out these values as they are the among best statistical indicators of overall talent/impact. By explicitly including these values, it allows us to compare the impact of overall player talent to the impact of the combination of specific player types and skills. This will prove crucial in analyzing the idea of team fit in the future.

**4.3.   Roster Representation**

Once we chose the 11 total factors to describe a player, the question became how to choose the players that represent a team in order to predict its SRS. There are numerous factors to consider here, including trades, signings and waivings, injuries, and a player's place in the normal rotation. First, we decided to prioritize players that had played the most minutes with a team, as these are the players that are most definitive of the team's identity. While this means a player who was traded mid-season could be used to represent their former team in the training data, the player's inclusion is necessary as the team's SRS was heavily influenced by their minutes. Additionally, we decided to include only the top 10 players from this group to define the overall roster. We reached this conclusion due to the average size of NBA rotations, which was estimated to be under 13.5 players for the regular season and under 9.4 players for the postseason (Fromal). As such, the final representation of a roster for any given year was a vector of length 113 (11 factors for 10 players, the team name, the team year, and the team SRS). To standardize the representation for improved learning, the player summaries were sorted by VORP in descending order.

**4.4.   Model Choice**

Initially, we attempted to build the STAR model using a neural network, given their history of success in identifying hidden relationships between input variables. However, we immediately ran into issues, as neural networks are incapable of robustly handling categorical data such as the player archetype. This led us towards a more robust model that could handle both categorical and numeric data: decision trees. Decision trees offer similar strengths to neural networks, as they are easily capable of representing relationships within the input data. Additionally, they are capable of very high accuracy scores, as each node of the tree is built by selecting the attribute and cutoff that maximizes the information gain for the resulting branches. However, decision trees are often prone to overfitting to the training data, especially when they are not pruned after training, leading to a model that is not capable of generalizing to unseen data. In order to combat this issue, we made use of a Random Forest classifier, a common machine learning bagging algorithm that builds many separate decision trees on subsets of the training data. Anytime a new example is classified, it is presented to each of the decision trees in the forest, and the result is a voted average of the resulting decisions. The idea behind bagging algorithms such as random forests is to reduce variance in the model by providing accurate input from many different sources. While the new model will not necessarily reduce bias (i.e., increase accuracy on the training data), it reduces the tendency of the original single decision tree to overfit.

**4.5.   Model Training**

The random forest algorithm was provided with a dataset composed of team roster vectors: one vector for each NBA team year dating back to 1985.

**4.6.   Results**

**4.6.1.  Training on all data up to 2019**

Our very first method of testing the effectiveness of the STAR model was to train on all the data we had available up to but not including the 2019-2020 season. We then tested on the single unseen year, as this is likely to be the most common method of practical application for the model. As the model predicts SRS, a numerical value, we can evaluate the model's performance by analyzing the correlation between the predicted SRS values and true SRS values. The R2 value when STAR was tested on the training data was 0.9854, an incredibly strong correlation. The resulting scatter plot can be seen in Figure 14.

**Figure 14:** *Predicted SRS of Training Data vs. True SRS of Training Data*

**Chart, scatter chart

Description automatically generated**

The predicted SRS values for the validation data (teams from 2019-20) had an R2 of 0.9429 with the true SRS values, also a very strong correlation. The resulting scatterplot can be seen below in Figure 15. Note that the predicted SRS in this set are all scaled down due to the reduced season length in 2019 caused by the Coronavirus pandemic.

**Figure 15:** *Predicted 2019 SRS vs True 2019 SRS*

**Chart, scatter chart

Description automatically generated**

**4.6.2.  K-Fold Cross Validation**

Next, we performed K-Fold Cross Validation to ensure that the model was able to generalize well to different sets of unseen data. After choosing K to be 6 (typical K values fall in the range 5-10, 6 allowed for data to be evenly split), the original dataset was separated into 6 random subsets. We then performed training and validation 6 separate times, each time withholding one subset from the training data in order to be used for validation. Over the 6 folds, the average training R2 was 0.98, and the average validation R2 was 0.772, still showing very strong correlation. The scatterplot of one of the six folds can be seen below in Figure 16. Due to the use of the random forest algorithm, we can reasonably assume the difference in training and validation scores is not due to overfitting. While no further testing was performed, we have hypothesized that this difference was observed due to training and testing on such a wide time period, in which the game of basketball has evolved significantly. The style of play in 1985 is vastly different from that of the early 2000s, which is vastly different from today's game. As such, there may be hidden factors regarding the evolution of the game that have not been accounted for within the model.

**Figure 16:** *Cross Validation Testing Scatterplot*

**Chart, scatter chart

Description automatically generated**

**5. Ideal Team Building Principles**

**5.1 Introduction**

Now that the STAR model has been built and confirmed to predict SRS accurately, we can use it to analyze a few common team building strategies in order to provide statistical insight into the ideal principles for NBA roster construction.

**5.2 The importance of depth**

All offseason, the Brooklyn Nets have reportedly been in trade talks to acquire James Harden, Bradley Beal, or some other third star to pair with Kevin Durant and Kyrie Irving. While many are excited by the idea of stacking star power, there are many fans and analysts who believe the Nets' current depth pieces to be more important. This begs the question: which is more valuable, star power or depth? Would a team be better off with multiple max-contract players surrounded by veterans on minimum deals, or a combination of multiple sub-All-Stars and solid starter level players? By making use of STAR, we can attempt to provide insight into the question.

In order to analyze the importance of depth, we made use of the Increase in Node Purity metric (INC). Essentially, INC scores measure how much the purity of resulting tree branches increases after splitting on an attribute, where purity is a measure of similarity between remaining examples. In essence, INC scores serve as a proxy for how important an attribute is in distinguishing varying levels of team strength. As such, we can analyze the importance of depth by plotting the INC scores of each of the 10 player VORP values on a roster, sorted in descending order. As seen in Figure 17, the top four players on a team (ranked by VORP) had INC scores over 2400, while the fifth player's VORP INC was only 1537, the sixth 895, and so on with continued decline in importance. As such, we can establish a basic heuristic: a team's core is made up of its four best players. While a team's starters are made of five players, and the sixth man is often a glorified role, the true differentiation of a team's strength comes from their top four players, with the third and fourth best player being particularly important to overall team success.

While it may seem counterintuitive that the third and fourth best players are more important than the first and second best players, a classic example can be seen in the recent dynastic Golden State Warriors. While unanimous MVP Stephen Curry was clearly the team's best player, there are countless YouTube videos, and sports articles titled something along the lines of "3 Reasons Draymond Green is the Warriors' Most Important player" or "Warriors' most important player? Game 3 shows why it might be Klay Thompson" (Jenkins, Robles). Clearly, neither player could lead the Warriors to a title without Stephen Curry. However, the counterpoint is that Steph might not be able to win a title without them either. While the contributions of the best or second-best player are often expected, foregone conclusions, the abilities of the third and fourth best players allow a team to separate themselves from the pack. Klay Thompson is not Stephen Curry, but he is an excellent player when compared to other players who may slot in at the third position on their team.

**Figure 17:** *INC Plot*

Chart, scatter chart

Description automatically generated

**5.3 Talent vs. Fit**

The question of talent versus fit has circled over NBA front offices for years. Is a team better off stacking as many talented players as it can, regardless of potential overlap in skills? Or is there an advantage to collecting synergistic pieces which can enhance one another's value with complementary attributes? Once again, by analyzing the Inc Node Purity scores for STAR's input variables, we can glean some insight into a potential answer.

As noted in section 5.2, the INC scores of the VORP values for a team's top four players are all greater than 2400. However, save for one variable, the features for all of the archetypes and 5 Factor data (our representation of the fit between player skills) had INC scores of less than 105. In comparison, the INC score of the 10th player's VORP sat at 356, still more significant than any of the classification/summary factors. In fact, the only non VORP/BPM factor with a score over 105 was the Offensive Talent Factor for the team's best player with an INC score of 712. Once again, this information allows us to create a basic heuristic: talent outweighs fit at all levels of the team. The only fit that matters to a significant degree is the ability of your best player to be an elite offensive option (teams led by defensive players have historically struggled to climb the standings to elite levels).

While many may reference the 76ers as a prime example of talent's inability to overcome fit, perhaps the issue is more complex than this simplification suggests. While the 76ers fit issues and lack of shooting/spacing have clearly compounded the team's weaknesses, it may also be true that their talent level is not as high as it is perceived to be. Over the last three years, the 76ers have consistently finished between the third and sixth seed, losing in the second round twice before losing in the first round this year. While this performance has been classified as a disappointment by media and fans alike, the results are more in line with expectations than many believe. Over the last five seasons, no team has made the Finals without a player of at least 5.4 BPM (Jimmy Butler, 2019 Heat). Even this threshold is on the low side, as the next lowest BPM cutoffs were at 6.6 (Stephen Curry, 2018) and 7.2 (Kawhi Leonard, 2018). Meanwhile, while the 76ers have two top 20 players in Joel Embiid and Ben Simmons, neither has a BPM near these cutoffs, with Embiid falling at 4.7 this year and Simmons at 3.6. While there may be significant issues with the fit between the two stars, perhaps this is just yet another example of the dominant importance of talent, especially elite talent, when predicting NBA success.

**5.4 Most Important Factors**

While it appears that talent overrides fit in most instances, the distributions of a player's 5 Factors still play an important role in the prediction of a team's SRS. By analyzing the INC scores for the 5 Factors summaries of each of the ten players on a roster, we can learn a few more interesting concepts. These concepts are visualized below in Figure 18 (the top VORP player was left out of the figure due to its incredibly high INC scores, which distorted the chart).

**Figure 18:** *INC Node Purity Scores for the 5 Factors*

Chart, line chart

Description automatically generated

While the Interior Focus, Perimeter Defense, and Unicorn attributes all seem to remain relatively low and constant for all 10 players, there are a few interesting patterns with Offensive Talent and Inefficiency. First, the importance of offensive talent drops off significantly from the second-best player to the third and drops again less significantly from the third to the fourth. Similar to the heuristic developed earlier regarding the importance of a team's top four players, this indicates that the core of a team's offensive abilities is derived through the scoring and playmaking abilities of its best three players (more emphasis on the top two), with only minor contributions from the rest of the roster.

The opposite trend appears to be true with the Inefficiency metric. While the INC scores for more highly rated players' Inefficiency ratings begin on the low end, they slowly grow in importance deeper into the roster. Intuitively, this makes sense given our prior knowledge, as bench role players are often dependent on the Offensive Engines to create open shots for them. At that point, the only value they can derive is by being as efficient as possible on these generated shots and not turning the ball over unnecessarily. This is very similar to the Lebron-style roster building from the last decade: surround Lebron James with as many shooter role players as possible to take advantage of the open shots he generates for teammates. The more efficient these shooter role players, the higher the rating of the offense, while inefficient role players can prove to be a detriment to the success of the team.

**6. Team Building Example**

**6.1 Introduction**

Now that we have established some basic heuristics regarding optimal roster construction, we can put them into action in order to build and test a theoretical team. Given the prevalence of the Philadelphia 76ers in this discussion, we decided to build a new team around 76ers star Joel Embiid, attempting to alleviate previously mentioned concerns regarding fit and talent levels on the team. We are assuming that all players in the NBA are available. Of course, any exercise of this nature must be bounded within the realm of realism, so we set a few constraints on the players available to our roster. First, the team's total salary had to remain under the NBA Luxury Tax for the upcoming season, 132.6 million dollars. Second, we had to set limits on the quality of players a team could realistically acquire. To set hard boundaries, we calculated the average BPM of each of the 10 players on strong teams throughout history (SRS >= 2) and set these values as the maximum cutoffs for player BPM values on our team. For example, our second-best player had to have a BPM <= 3.4, and the third best player needed a BPM <= 2.1. The results of the roster construction exercise can be seen below in Figure 19.

**6.2 Roster Building**

After starting our team with Joel Embiid, it was time to choose a co-star. However, with a BPM cutoff of 3.4, our options were limited. A few players were considered, including Kyle Lowry, Bam Adebayo, and Domantas Sabonis. However, the ultimate selection was Bradley Beal of the Washington Wizards, as the most important attribute for a team's second-best player is their Offensive Talent, a metric in which Beal dominated the competition. By combining this with our prior knowledge of Beal as an elite scoring guard with a strong handle, the fit with our dominant center seemed to make for a strong pairing.

The choice for the third player made use of a couple of the previously discovered heuristics. First, the idea of top end talent (the top 3-4 players) outweighing depth was taken into consideration. The question of whether to add a third player on a max contract was a valid concern when building the team. Would a third expensive piece make the rest of the roster suffer? However, as noted before, STAR believes more in top end talent impacting overall team success than depth pieces, leading us to acquire a third player on a max contract: Cleveland Cavaliers Forward Kevin Love. Love was chosen over players of similar level BPMs such as Brandon Ingram and Pascal Siakam due to his incredible combination of high Scoring Talent (1.2, 88th%) and low Inefficiency (-1.1, 14th%), both of which are very important for a team's third best player. The low Inefficiency was particularly important, as, in theory, it allows him to fit more seamlessly with our two high volume scorers, playing a complementary scoring role without turning the ball over or missing shots at a high rate.

According to Figure 18, the most important attribute for a team's fourth best player is their Inefficiency, leading us to believe this spot would best be used on an extremely efficient complementary piece. While talented options like Buddy Hield, Jamal Murray, and Ricky Rubio fit under the BPM threshold, they were either too expensive or too inefficient to fit the criteria in this section. Ultimately, the choice came down to Seth Curry and Justin Holiday due to their incredible Inefficiency scores, with Holiday winning out due to an advantage in BPM and Perimeter Defense.

The final piece of the starting five allowed for more flexibility, as none of the 5 Factors appeared to be dominantly important for this position. As such, we were able to evaluate the current strengths and weaknesses of the team and noticed we were severely lacking in players with strong Perimeter Defense and Unicorn grades. Fortunately, Lonzo Ball filled the void nicely with 81st percentile Perimeter Defense and 97th percentile Unicorn grades while falling right under the BPM cutoff.

The rest of the roster was composed using similar techniques, limiting BPM each time and emphasizing certain attributes given the data in Figure 18. Each of Elfrid Payton, Duncan Robinson, Bobby Portis, Josh Hart, and Meyers Leonard offered a combination of skills which STAR viewed in high regard, all without exceeding the Luxury Tax threshold.

**Figure 19:** *Player Profiles for Rebuilt 76ers Team*

A picture containing chart

Description automatically generated

Chart

Description automatically generated

A picture containing chart

Description automatically generated

Chart, bar chart

Description automatically generated

A picture containing timeline

Description automatically generated

Chart

Description automatically generated

A picture containing table

Description automatically generated

**6.3. Results**

Finally, with the rebuilt 76ers roster complete, we entered the newly constructed team into STAR in order to predict the strength of the team through their estimated SRS. While the original 76ers finished with an SRS of 2.22 this season, STAR predicted the newly constructed 76ers to finish the season with an SRS of 4.37, which would rank 7th in the league. This improvement appears in-line with our expectations, as we were able to address the fit issues in Philadelphia by combining pieces with STAR-approved complementary skill sets. However, as we acknowledged earlier, the 76ers still struggle from the lack of a dominant top 5 player with a BPM in the 6+ range. While the team appears to be markedly improved, they still fall short of the true title contenders in terms of SRS. While building a team with proper fit can certainly help to alleviate some issues, at the end of the day, there is no replacement for high-end talent.

**7. Conclusion**

In building the STAR model, we show that it is possible to accurately predict the quality of a team (SRS) using machine learning techniques. The process of building the model also yielded many insights. We found that summarizing the skills of NBA players through clustering and compression of their advanced and per game statistics (both counts and rates) created accurate representations of a player's role and on-court contributions. Our methods give NBA front offices the ability to not only quantitatively evaluate their team but also to better understand their players' on-court skills, regardless of position.

Additionally, we used STAR to answer many of the big questions that front offices face when attempting to construct the perfect roster. Ideological concepts such as the importance of depth or the benefits of talent versus fit were examined closely in order to deepen our understanding of the factors behind a successful roster. Analysis of the model's results show that a team's core is made up of its 4 best players. While deeper rotation players can still provide beneficial impact, the true differentiators lie at the top end of the roster. As for the debate between talent and fit, we discovered that talent/impact appears to trump fit/role. Finally, INC score analysis was used to determine which factors among a team's top ten players were the most impactful, with Offensive Talent and Efficiency ranking as the top two in most scenarios.

Like with all models, there is room for improvement within STAR. A couple of the drawbacks stem from the problem that we cannot accurately predict future statistics, and in particular, the development of young players. Right now, STAR predicts a team's future success based on its players' past season statistics and cannot handle rookies as they do not have past statistics. In the future, inputting accurately predicted future statistics can improve the model's ability to look forward rather than just analyze the past and present. On a similar note, STAR could be adjusted and improved to serve as more of a real-time rating system by emphasizing recent data (weight the last 10 games more heavily) rather than weighting all the data from a season evenly. This would allow the model to stay current and adjust to new, relevant information with higher urgency. Additionally, at its current stage, STAR is not a good indicator of playoff success, as it is trained on regular season SRS data. While SRS is highly indicative of overall team strength, playoff basketball is a different game than regular season basketball, and regular season findings do not always translate to playoff scenarios. A different STAR playoff model could be trained in the future with some form of a weighted playoff net rating as the predicted outcome rather than SRS. Finally, while we do not view it as a major issue, a majority of a team's success is weighted on its players' VORPs instead of other inputs like archetypes and the 5 Factors. This can lead to disproportionate influence from a few numbers in the model, especially when these numbers are not necessarily translatable from team to team. In the case of VORP, while it may serve as the best estimate of a player's overall impact, it is not resistant to change due to shifting surroundings. Just because Rudy Gobert has a VORP of 3.3 on the current iteration of the Jazz does not mean he would have the same VORP if traded to another team like the Celtics. However, given the lack of better alternatives available in the public domain, VORP serves as a relatively accurate indicator of player strength. While this high influence of a few factors may not be ideal, we believe a team's talent is the most important factor in determining success, and the model is simply reflecting that concept by heavily weighting VORP.

STAR's capability to build teams from scratch and then predict their SRS provides more room for insight and exploration in the field of roster management. STAR allows teams and front offices to experiment with adding and removing specific players and their corresponding statistical footprints without consequence. This customizability is especially useful to NBA front offices in free agency and trade decisions, as they can see the concrete change in predicted team performance due to a roster move. The ability to explore player and team combinations in an unbounded manner opens doors to creative new roster construction methods, less reliant on subjective beliefs and instead backed by machine learning and historical trends.

**References**

[1] Fromal, Adam. "NBA Metrics 101: Which NBA Playoff Myths Are Actually True?" *Bleacher Report.* 9 May 2017.

[2] Kalman, Samuel, and Bosch, Jonathan. "NBA Lineup Analysis on Clustered Player Tendencies: A New Approach to the Positions of Basketball & Modeling Lineup Efficiency of Soft Lineup Aggregates." *MIT Sloan Sports Analytics Conference.* 2019.

[3] Basketball-Reference. "2019-20 NBA Player Stats: Per Game." *Sports Reference LLC.* 2019.

[4] Basketball-Reference. "2019-20 NBA Player Stats: Advanced." *Sports Reference LLC.* 2019.

[5] Basketball-Reference. "2019-20 NBA Season Summary." *Sports Reference LLC.* 2019.

[6] Basketball-Reference. “2019-20 NBA Preseason Odds.” *Sports Reference LLC.* 2019.

[7] Basketball-Reference. *Sports Reference LLC.* 1985-2020.

[8] "NBA Player Salaries." *HoopsHype.* 2020.

[9] *NBA.com.* 2020.

[10] Robles, Miguel. "3 Reasons Draymond Green is the Warriors' Most Important Player." *12Up.* 20 October 2017.

[11] Jenkins, Bruce. "Warriors' most important player? Game 3 shows why it might be Klay Thompson." *San Francisco Chronicle.* 6 June 2019.