Using Machine Learning to Construct Optimal Team Rosters in the Modern NBA

Jaden Patel

St. Paul's School

Abstract

This study analyzes NBA roster composition over a 10-year period (2014/15 to 2023/24), aiming to identify optimal player archetypes and positional balances for maximizing team success. Detailed individual and collective performance and physical trait data from 300 distinct teams and 3557 players was used. Players were clustered into ten archetypes and three general positions (Guards, Wings, and Bigs) through k-means clustering. A supervised learning (gradient boosting) model was then employed to predict team win totals based on archetype and position profiles. Results highlight the critical role of 3-point Specialists and Defensive Wings in modern NBA success, underscoring the value of versatile, low cost players - role players who contribute on both ends of the floor. Rosters with balance across positions and archetypes yielded the highest predicted win totals, while over-reliance on high-usage offense roles such as Initiating Guards and high-cost Playmaking Scorers led to diminishing returns. These findings provide

actionable insights for NBA front offices seeking to optimize roster construction within the context of evolving league dynamics and salary cap constraints.

Introduction

In the dynamic landscape of the modern NBA, a large part of team success relies on the realized synergies across players on the roster. Basketball is a team sport, so although the individual talent of the players is a large contributor to winning, the *types* of players and the *roles* they play on the team also hold great importance. NBA front offices have the task of constructing a roster of the best possible players that also can work together as a team to play winning basketball. In an era of “super-teams” and the continual movement of superstar players, it is crucial that teams deploy a roster of the right mix of players from top to bottom to build a winning team.

The Golden State Warriors are the most recent NBA dynasty, winning four championships in an eight year span (2015, 2017, 2018, 2022). The Warriors' success is a testament to the importance of roster construction. While the superstar talents of players like Stephen Curry, Klay Thompson, and Kevin Durant were instrumental, their dominance was also fueled by the contributions of role players such as Draymond Green, Andre Iguodala, and Andrew Wiggins, each delivering a specific contribution to the team. This balance between star power and complementary pieces has underscored the value of building a team that performs as more than the sum of its parts. After losing to the Toronto Raptors in the 2019 NBA Finals, Kevin Durant decided to team up with Kyrie Irving on the Brooklyn Nets, which shortly thereafter also added James Harden. This new superteam with three of the league’s top offensive talents had the expectation of becoming the next dynasty. However, they failed miserably, not even making a conference finals. A combination of injuries, poor fit, and a shallow roster lacking key depth-pieces led to the Nets underwhelming performance.

Since 2019, a different team has won the championship each year, illustrating the growing importance of adaptability, balance, and strategic roster construction in an increasingly competitive league. Many of these teams built rosters centered on a homegrown superstar player around whom they were able to develop and select the right supporting cast that could bring home a championship. The 2021 Milwaukee Bucks succeeded with a homegrown core led by Giannis Antetokounmpo and carefully selected complementary pieces. The Denver Nuggets' 2023 title emphasized the value of continuity, chemistry, and maximizing the skill sets of versatile players around their superstar center Nikola Jokic. The Nuggets developed a core centered around both Jokic and star point guard Jamal Murray that was finally able to get it done in 2023 with a well developed and holistically constructed roster.

There have been a number of recent studies that analyzed lineup combinations. (Riccardi, 2024) grouped players into six clusters and used linear regression models and measurements predicting teams' pythagorean win percentage considering the total minutes played for the season by each team's best player, in order to quantify the importance of each play-style cluster to winning games. This study focused on complementary player types, analyzing what player-types best fit around a team's star player of a given archetype. Kalman and Bosch (2020) modelled lineup efficiency using players from 2009-2018 by clustering players into nine archetypes and then created five-man lineup combinations with those nine clusters. They used linear regression to determine the impact of each of the clusters on the Net Rating of their five-man lineup, and a Random Forest Model was used to predict the Net Rating of all possible lineup combinations. This analysis examines the on-court performance of different lineup combinations.

The goal of the following analysis is to cluster players into different archetype groups that represent their play-style on the court and role on their team, and use the composition of past NBA teams to determine the ideal roster makeup using these archetypes. Rather than looking at the on-court effectiveness of different lineup makeups as others have done, I broaden this analysis to the team's roster as a whole across the entire season to see what types of rosters win the most. I use data from the 2014/15 to 2023/24 seasons in order to analyze the relationships across player archetype combinations and team success variables.

Methodology

Variable Selection

I collected detailed performance and physical data for every player in the NBA from the 2014/15 to 2023/24 seasons. Data was collected from *Basketball Reference* and *Kaggle*. It was then compiled into one dataset that contained ten years of data. This ten year span provided an optimal balance between recency and sample size. Sixteen different variables were used, which were meant to encapsulate player tendencies and their physical profiles, which influence their classified role. 'Per 100 Possessions' statistics and shot distribution statistics were used for the on-court performance data. Basketball Reference provides in-depth shooting statistics that are useful to analyze players' offensive profiles. A player's shooting statistics are represented by the variables short, post, and mid. These are the shot attempts per 100 possessions a player has from those ranges. This is calculated as the percentage of total field goals a player attempted from 0-3 feet, 3-10 feet, and 10 feet to 3-point range, multiplied by Field Goal Attempts (shots taken) per

100 possessions. These modifications were made for distribution statistics in order to adjust for sample size. Additionally, a “3” represents the percentage of a player’s 3-point attempts to be assisted, multiplied by their total 3 point attempts per 100 possessions. This new calculated variable is the total number of assisted 3 pointers per 100 possessions, which now takes into account how often the player actually shoots ‘3s.’ The same method was applied for a “2,” but for2-point attempts.

Filtering

The data was filtered to only include players who played at least fifteen games and averaged at least fifteen minutes per game in a given season. This ensured there was a large enough sample size of performance data to accurately classify the player into an archetype group.

There were some centers who had a 3-point% of 1.0 because they only had one 3-point attempt, so they went 1/1 = 100% or 1.0. Because of this, players with a 3-point% of 1 were changed to 0. This presents a more accurate representation of their 3-point shooting ability, because even though they made the single three pointer they attempted, this is simply due to random chance and the fact they did not attempt any more threes shows they are not meaningful contributors as three-point shooters.

Clustering

To classify NBA players into meaningful archetypes, I utilized the k-means clustering algorithm, a widely used machine learning method for unsupervised data analysis. Before applying k-means clustering, the dataset was standardized to ensure all variables were on the same scale. Standardization was necessary because raw statistics, such as assists/100 possessions and field goal attempts, exist on different scales and could disproportionately influence the clustering process. Each metric was normalized using z scores. A dataset of z scores will always have a mean of 0 and standard deviation of 1. This standardization allows us to compare different statistics based on how that player did relative to every other player in the dataset. The k-means algorithm was selected for its simplicity and interpretability. After testing a variety of different cluster quantities, I decided to specify ten clusters, representing the number of player archetypes that best described the diversity of roles in the modern NBA. The algorithm assigns players to clusters by minimizing the within-cluster sum of squared distances (WCSS) between each player and the centroid of their cluster. It iteratively refined cluster centroids until convergence, ensuring that players with similar statistical profiles were grouped together. The output of the clustering process yielded ten distinct player archetypes, each representing a unique style of play. For example, some clusters captured scoring and playmaking abilities (Playmaking Scorers), while others reflected defensive dominance (defensive anchors) or specialized skills like three-point shooting (sharpshooters). Each player was assigned to the archetype with a centroid closest to their statistical profile, forming the foundation for analyzing roster compositions and team success.

Figure 1: K-means archetype cluster centers for each standardized attribute

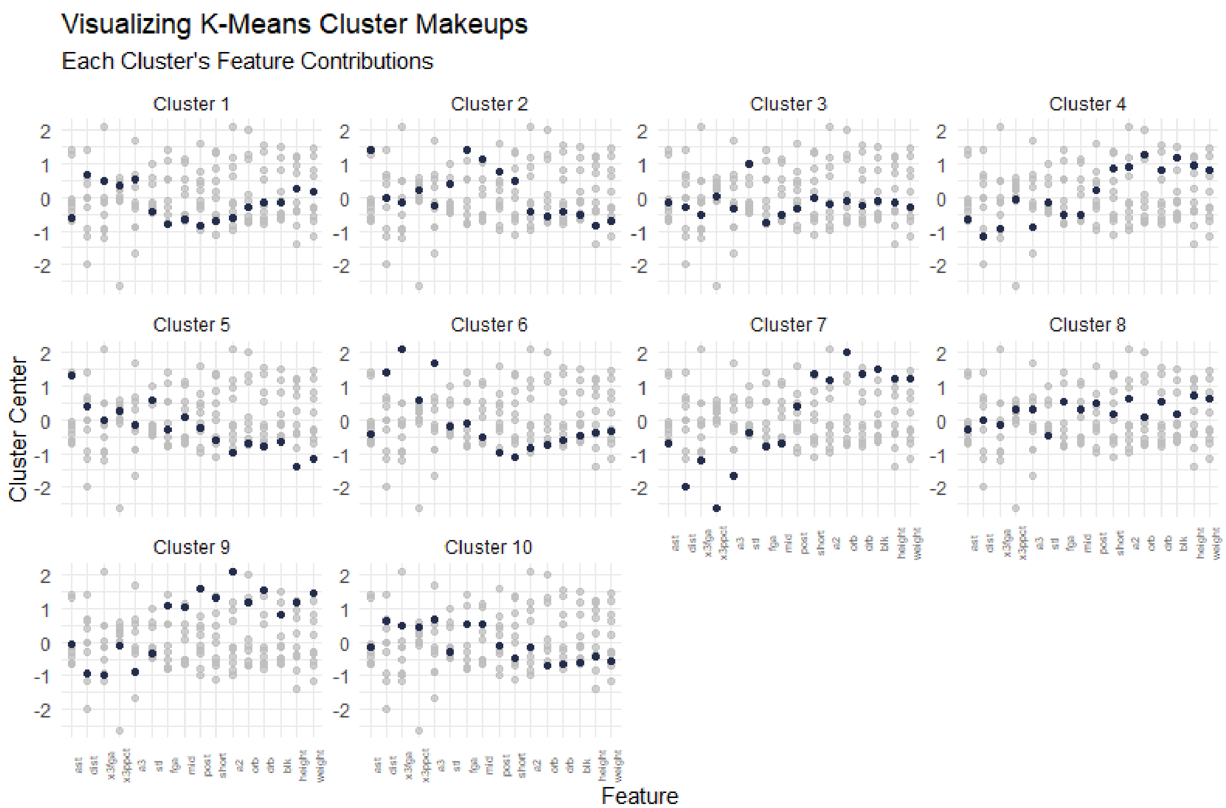


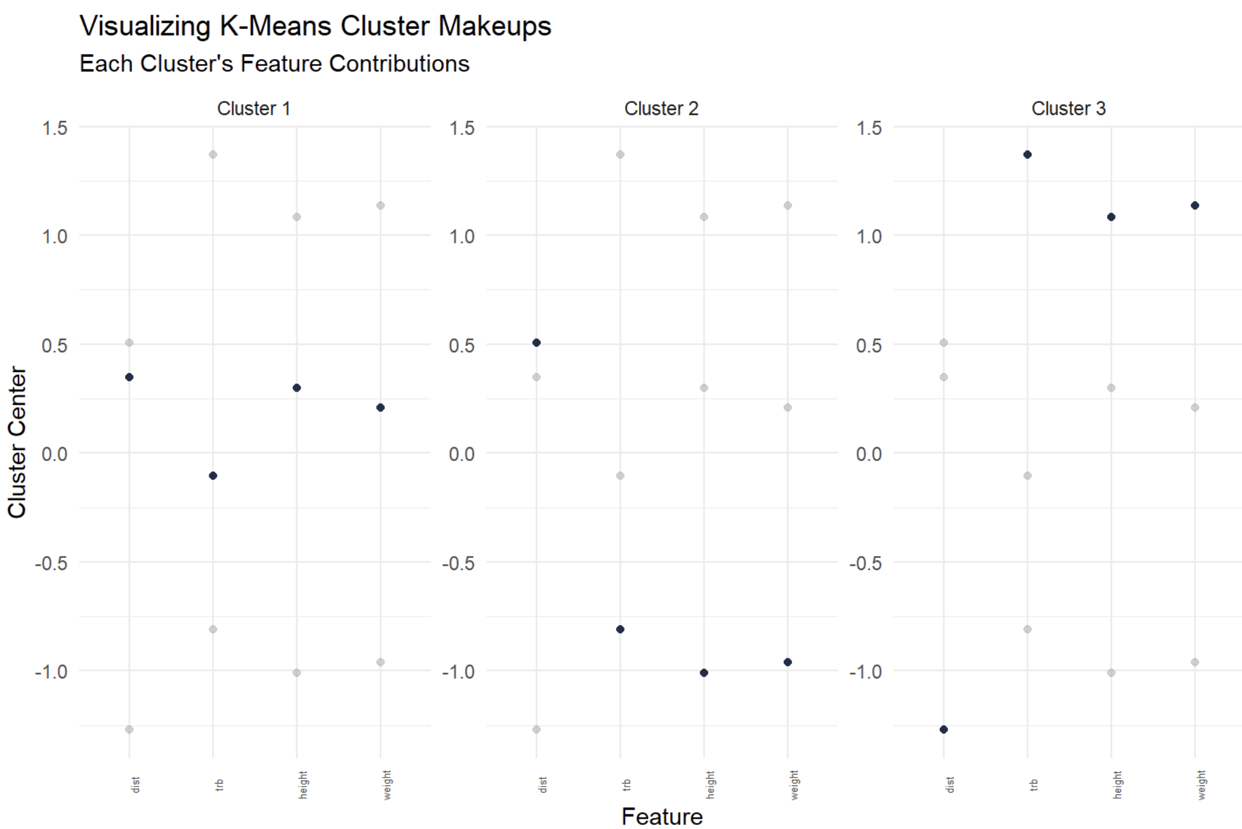
Table 1: Archetype descriptions

Archetype	Description	High Stats	Low Stats	Player Examples
3pt Specialist	Catch-and-shoot 3 point shooter. Smaller complementary players who get minutes because of their three point shooting ability	3pt Attempt Rate 3pt% Assisted 3s	0-3 Ft Attempts 3-10 Ft Attempts	‘23 Duncan Robinson ‘20 Kyle Korver
Defensive Wing	High-end lockdown perimeter defenders. Non-shooters who can be good playmakers and do most of their scoring by driving to the rim.	Steals Assists	FGA 10ft-3pt Scoring	‘21 Matisse Thybulle ‘22 Gary Payton ll
Initiating Guard	Small pass-first ball handler who plays around the perimeter and is typically a good 3pt shooter.	Assists 3pt Attempts	Height Weight Rebounds Blocks	‘23 Andrew Nembhard ‘18 Rajon Rondo
Playmaking Scorer	High volume scorer that is effective with the ball in his hands.	FG Attempts Assists	Blocks Rebounds Height Weight	‘21 Tyrese Maxey ‘20 Kyrie Irving ‘24 Luka Doncic
Ranged Scorer	Well-rounded scorer who is high in volume but typically not as much of a playmaker as a	3pt Shooting Assisted 3s	Blocks Rebounds	‘18 Klay Thompson ‘22 Anfernee Simons

	Playmaking Scorer. Shoots a lot of threes but also has an effective mid range jumper and can drive to the basket.	10ft-3pt Attempts		
Rim Protecting Big	Traditional stay-at-home, defensive-minded bigs. Their role on the team is to use their size to block shots and grab rebounds.	Blocks Rebounds Short Range Height Weight	3pt Shooting Assists	‘15 Andre Drummond ‘18 Hassan Whiteside
Scoring Big	High volume offensive big men who are dominant in the paint and also have a strong mid-range and post game. They are also able to rebound and block shots due to their size.	Short/Post/Mid Range Attempts Blocks Rebounds Height Weight	Steals Assisted 3s	‘18 Giannis Antetopounmpo ‘19 Joel Embiid
Shooting Forward	Effective 3pt shooters who are low-volume scorers and are not good ball-handlers and playmakers.	3pt Shooting	Assists FGA Short/Post/Mid Range Attempts	‘24 Nicolas Batum ‘20 Royce O’Neal
Stretch Forward	Larger wings and athletic big men who can stretch the floor with their shooting ability. Bigger version of a Ranged Scorer who are also effective paint players and can rebound, but are not as dominant inside as Scoring Bigs. Big, well rounded offensive players who can also contribute defensively.	All Scoring Height	Steals	‘23 Lauri Markannen ‘22 John Collins
Versatile Defensive Big	Athletic power forwards and centers who can guard multiple positions and have more range than traditional rim protectors.	Short Range Attempts Height DReb	Assists 3pt Shooting	‘24 Aaron Gordon ‘21 Brandon Clarke

Additionally, players were clustered into one of the three different positions - guard, wing, and big. These position groupings focus on players’ physical traits and general role on the court rather than specific roles. Mean shooting distance, total rebounds, height, and weight were used as the four parameters. Guards will be small and shoot further out and will be low in rebounds due to their playing around the perimeter. These are more traditional point guards and some shooting guards. Bigs will be tall and heavy with most of their shots being close to the basket, and they also get a lot of rebounds. Wings have a wider range than guards and bigs for these attributes. Some are 3pt shooters whereas others score from driving inside. Ball handling /Shooting Wings can play as guards in a big-man lineup, and bigger wings can play as a big in a small ball lineup. Their main job is not to grab rebounds, but they spend more time around the paint and are typically bigger than guards, so their rebound counts are higher. Grouping players into these three generic positions will be useful for the model to build a balanced lineup.

Figure 2: K-means position cluster centers for each standardized attribute



Roster Analysis

After denoting an archetype to each player, each team's roster was transformed into a profile based solely on player archetypes. Individual players were replaced by their corresponding archetype designation, resulting in team-level archetype compositions. For instance, if a team had two Playmaking Scorers, one Defensive Anchor, and one Stretch

Forward, the team’s archetype profile reflected these counts. A games-played percentage (GP%) for each player was calculated by dividing the players' games played by the total number of games their team played that season, which can be found by adding their wins and losses. This statistic represents the percentage of total possible games that were played by the player. For example, a player with 82 games played in an 82 game season would have a GP% of 1, while a player who only played 41 games in an 82 game season would have a GP% of 0.5. This number helps reflect how often the player was actually available to play for their team. This adjusts for players who got injured or joined/left the team partway through the season. These players will naturally have less of an expected impact on their team’s wins because they were not on the active roster the entire season.

$$Games\ Played\ \% = \frac{Games\ Played}{TeamWins + TeamLosses}$$

Win Contribution enumerates the number of wins toward which a player contributed to their team based on their share of the team’s total minutes played that season. The player’s minutes played is adjusted so that it reflects an 82 game season where each game is 48 minutes long. This adjusts for shortened seasons and additional minutes played due to overtime games. This adjusted minutes played number is divided by the total possible minutes allocated by a team in a season. The 82 games played in the numerator and denominator cancel out, so we divide by 240 minutes, which is simply 48 minutes multiplied by five players on the court, multiplied by the total games played by the player’s team that season to account for shortened seasons. This differs from GP% because it uses minutes played to estimate the player’s contribution to the team’s win total, rather than simply availability.

$$Win\ Contribution = \frac{MinutesPlayed \times Wins}{240 \times (Wins + Losses)}$$

Team Adjusted Wins is the sum of every player's Win Contribution to the team. This new wins total only takes into account the amount players in the dataset impacted the team's win total, leaving out the impact of players who did not meet the criteria to be in the original dataset.

To understand the impact of specific archetype compositions on team success, teams were grouped based on the count of each archetype. For example, all teams with exactly two Playmaking Scorers were grouped, and the average win percentage for these teams was calculated. This analysis was performed for all possible counts of each archetype. Values represented the average win percentage for all teams with the given archetype count. This table served as a foundational dataset for evaluating how different archetypes influenced team success. Larger archetype counts had small sample sizes, which led to a lot of variation between the mean win percentage of different archetypes with that given count. As the counts increase, the variation generally increases as well, which makes sense because it is more likely that a team has only one player of a given archetype than four. Therefore, only counts of three or less were used for this analysis because the values for larger counts are unreliable, and it is reasonable to assume that having four or more of the same type of player would not be beneficial for a team.

Figure 3: Mean adjusted wins for each distinct archetype + count combination

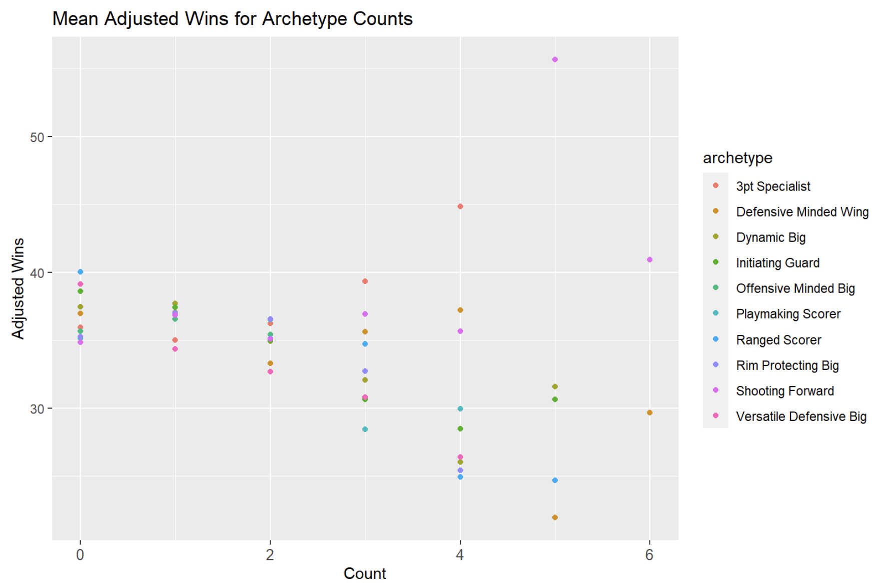
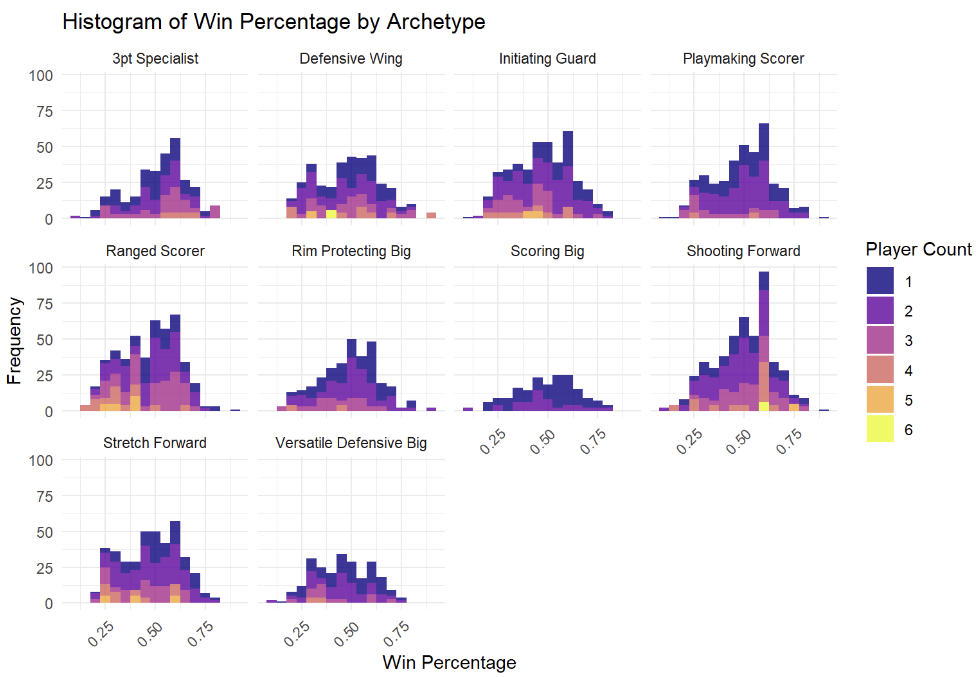
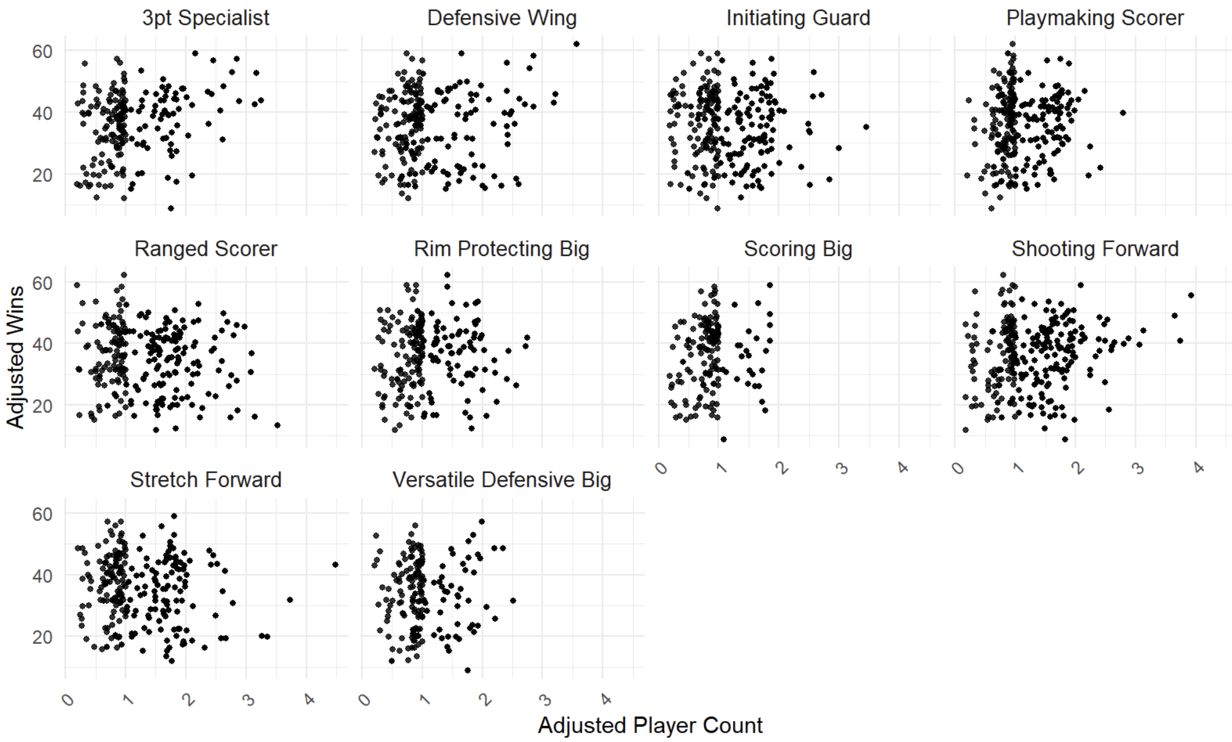


Figure 4: Win percentage distribution for each archetype



Rather than each player being worth 1.0, their value was instead their games-played (GP%). Rather than a team's total player count in a season being the raw player count, it was adjusted to be the sum of each player's GP%. The total number of players for each archetype on each team was found by aggregating players' GP%. The same process was applied for positions. This allowed for the archetype and position composition of all 300 distinct teams from all ten seasons to be found. All 300 teams' data with their Adjusted Wins and adjusted player count for each archetype and position was compiled into one dataset. The graphs above show the effect of archetype count on wins, adjusted to exclude the impact of players who did not meet the criteria to be in the analysis. Each point represents a different team, so there are 300 total points on each graph. In order to calculate a team's position profile, I had to define each position. The position profile of each archetype was the percentage of players of that archetype that played each position. For example, 71% of 3pt Specialists were guards and 29% were wings, so one 3pt Specialist would equate to 0.71 guards and 0.29 wings. A team's count for each of the three positions was calculated by taking the sum of the position weight of that particular position for all archetypes on the roster.

Figure 5: Scatterplot of the effect of the adjusted player count on adjusted wins for each archetype



Gradient Boost Model

A Gradient Boosting model was implemented in Python using the XGBoost library to predict win totals for every possible roster combination. The training dataset consisted of archetype and position profiles for NBA teams from the 201415 to 2023/24 seasons, incorporating thirteen features: ten player archetypes and three position categories (Guards, Wings, Bigs). The inclusion of position as a feature ensured roster balance in the predictions. The model aimed to identify optimal archetype combinations while maintaining positional integrity. Gradient boosting is a machine learning technique designed to improve predictive accuracy by iteratively combining the outputs of multiple weak learners to create a strong, cohesive model. Weak learners, in this context, are models that perform only slightly better than random guessing, and are often represented as decision trees with shallow depths. A decision tree is a hierarchical structure where data is split into branches based on feature thresholds, ultimately leading to predictions at the leaf-nodes. These trees are intuitive and excel at capturing simple patterns in data. However, when used individually, their predictive power is limited due to their tendency to underfit complex datasets. The modeling process begins with an initial prediction, which in this case could be the mean of the target variable - i.e., wins. Residuals, or the differences between actual and predicted values, are then calculated to identify areas where the model underperforms. Subsequent decision trees are trained to predict these residuals, effectively modeling the errors of the prior iteration. Each tree contributes to the overall model by addressing these residuals, and its predictions are scaled by a learning rate, which determines the proportion of the residuals it corrects. This iterative process continues, with each tree refining the model's understanding of the data by focusing on the cases that are hardest to predict, resulting in a model that captures intricate, nonlinear relationships. A dataset of all possible roster combinations, constrained to a maximum of three players per archetype, was generated, resulting in 39,853 unique rosters. The trained gradient boost model predicted the win count for each roster.\

Results

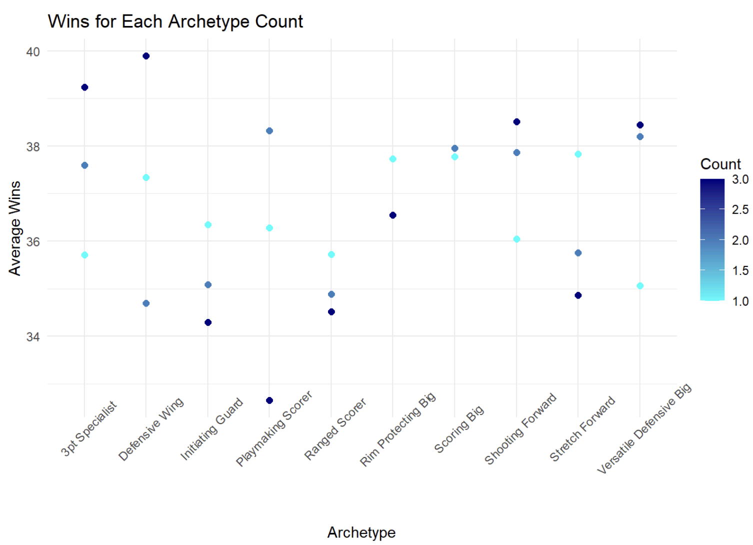
The gradient boost model made win predictions for all rosters in the dataset. Table 2 shows the roster with the highest predicted wins. This roster can be best summed up as a '3&D' team, with a focus on 3pt shooting and defence. In practice, this roster would likely be centered around one of the Scoring Bigs, who would be the primary scorer. The starting lineup would most likely consist of a 3pt Specialist and Defensive Wing as the guards, the Stretch Forward as the small forward, and the two Scoring Bigs at power forward and center. The remaining 3pt Specialists and Defensive Wings would be complimentary pieces coming off the bench, along with a Rim Protecting Big to provide strong interior defence. The most similar NBA teams are the 2024 Golden State Warriors, 2021 Golden State Warriors, and 2024 New Orleans Pelicans.

Table 2: Roster with highest predicted wins

3pt Specialist	Defensive Wing	Initiating Guard	Playmaking Scorer	Scoring Big	Shooting Forward	Stretch Forward	Versatile Defensive Big	Rim Protecting Big	Ranged Scorer	Guards	Wings	Bigs	Wins
3	3	0	0	2	0	1	0	1	0	3.39	4.45	3.12	52.06919

The analysis highlights the archetypes most strongly associated with higher win totals. Three Defensive Wings and Two Scoring Bigs and One Rim Protecting Big were consistently linked to the highest performing rosters. Conversely, archetypes such as Initiating Guards, Ranged Scorers, and Playmaking Scorer, while valuable in specific contexts, exhibited diminishing returns when overrepresented in roster construction.

Figure 6: Mean adjusted wins for each distinct archetype count in model predictions



According to the model, 3pt Specialists are particularly beneficial to have, shown by teams with increasing numbers of 3pt Specialists having positive impacts. This shows the importance of 3pt shooting in the NBA, with the continually increasing reliance on 3pt shooting. Although these 3pt Specialists typically are not very versatile and aren't offensive play drivers, these are valuable players to have coming off the bench with cheaper contracts. Additionally, we are seeing more players who are primarily catch-and-shoot 3pt shooters that can also provide some additional value besides three point shooting.

The highest average predicted wins were for rosters with three Defensive Wings, however the lowest predicted wins were rosters with two Defensive Wings, with one Defensive Wing being around average. This significant range makes it difficult to determine the true value of Defensive Wings, as there is no reliable trend. Shooting forwards on the other hand show an increase in wins for an increase in count. This shows a consistent upward trend giving more reliability to the number. Initiating Guards were overall the least favoured archetype by the model. According to the model, it is best to only have 1 Initiating Guard, but even those rosters have a below average number of average wins. The model also favoured Scoring Bigs over Playmaking Scorers. These two archetypes can be seen as the two primary option archetypes. Most team's offensive focal point will be either a Playmaking Scorer or Scoring Big. The model slightly prefers teams to have a big as their primary option rather than a guard. Having multiple players in higher usage roles where they are most effective with the ball in their hands shows diminishing returns - Initiating Guards and Ranged Scorers show a steady decline in average wins for each additional player of that role on the roster. Having 3 Playmaking Scorers on a roster is also not a good idea. A balanced composition is critical for maximizing wins. Teams with a moderate distribution of 3-4 Guards, 4-5 Wings, and 2-3 Bigs achieved the highest predicted win totals. Over-reliance on any single role group led to diminishing returns. This emphasizes the importance of versatility and adaptability in roster construction. Notably, teams with a surplus of wings consistently outperformed teams heavily skewed toward Guards or Bigs. This reflects the increasing demand for players who can excel in multiple roles, particularly on the defensive end, while maintaining offensive efficiency.

Discussion

The prominence of 3-point Specialists and Defensive Wings shows the critical role of players who can effectively stretch the floor on offense while making substantial defensive contributions. These archetypes highlight the value of elite complementary players who excel without requiring high offensive usage. Players such as OG Anunoby and Derrick White exemplify the importance and value of these types of players in the modern NBA. They provide offensive spacing through efficient 3-point shooting and are also high end defenders, making them indispensable assets in modern roster construction. Their ability to contribute significantly on both ends of the court without dominating possessions exemplifies the increasing demand for versatile, low-maintenance players in today's NBA. The value of these 3&D wings is recognized throughout the NBA, which has led to many of them getting bigger contracts. However, teams can find less versatile 3pt Specialists and Defensive Wings that come off the bench who provide good value for a lower cost.

The model revealed diminishing returns for certain archetypes when overrepresented, such as Initiating Guards, Ranged Scorers, and Playmaking Scorers, suggesting that offensive roles requiring high usage and ball dominance are best limited to one or two players per team. Conversely, archetypes like Scoring Bigs and Shooting Forwards demonstrated consistent value, reinforcing the need for a balanced approach that integrates inside scoring with floor spacing.

The model's preference for Scoring Bigs over Playmaking Scorers as primary options further supports the idea that a team's offensive focal point may benefit from a size advantage in the modern game.

The data also highlights the increasing importance of wings in roster construction. Teams with a surplus of versatile wings consistently outperformed those heavily skewed toward Guards or Bigs, reflecting the demand for multi-role players who can switch defensively and maintain offensive efficiency. The nuanced trends associated with archetypes, such as the mixed results for Defensive Wings and the steady rise in wins with additional Shooting Forwards, emphasize the need for adaptability and context-aware roster decisions.

Overall, the study demonstrates that while archetype-specific strengths are vital, success ultimately depends on balance and versatility. Teams that diversify roles, maintain a moderate distribution across Guards, Wings, and Bigs, and limit redundancy in high-usage positions are more likely to achieve optimal performance. These insights provide a robust framework for NBA front offices to evaluate and construct rosters that maximize their competitive potential.

Several limitations are present in this study. First, the clustering process required excluding players who did not meet the threshold of playing at least fifteen games and averaging fifteen minutes per game in a season. This exclusion reduced the dataset's comprehensiveness and potentially omitted impactful players, especially those affected by injuries, trades, or rotational changes. This limits the ability to construct a complete picture of team archetype compositions. Ideally, every player who appeared for a team would be included to create a more accurate representation of roster dynamics. This would also allow for a better understanding of how bench depth and lesser-used players contribute to team outcomes. Consequently, team win contributions had to be estimated based solely on the players included in the dataset.

Adjustments were made to account for personnel changes and injuries by treating players' contributions as proportional to their time on the court, but this approach is only an estimation.

Another limitation lies in the reliance on aggregated win totals as a measure of team success. Win totals reflect performance across an entire season, which inherently masks game-to-game variations and the situational impact of individual players. Treating players as the sum of their contributions rather than analyzing their roles and performance on a game-by-game basis simplifies the complexity of real-world basketball dynamics. This includes external factors such as coaching styles and the overall talent on the roster, which significantly influence roster construction and team performance. Future research could integrate these contextual elements to provide a more holistic view of what drives team success in the NBA.

These limitations highlight that the results of this study, while insightful, are far from providing a definitive answer to the complexities of NBA roster construction. The nuances of basketball, from player synergies to the impact of external factors, cannot be fully captured in a single model or dataset. However, this analysis offers a strong foundation for understanding the relationships among player archetypes, positional balance, and team success. By identifying key trends and archetypes that contribute to winning rosters, the findings serve as a starting point for more comprehensive explorations into optimal roster composition. Future research can build on this work by incorporating additional in-depth variables, exploring interaction effects, and creating more adjustments to capture the intricacies of team-building in professional basketball. This study lays the groundwork for a deeper understanding of how teams can maximize their potential through strategic roster construction.

References

Kalman S, Bosch J (2020) NBA lineup analysis on clustered player tendencies: a new approach to the positions of basketball & modeling lineup efficiency of soft lineup aggregates. In 14th annual MIT Sloan Sports Analytics Conference https://global-uploads.webflow.com/5f1af76ed86d6771ad48324b/5f6a65517f9440891b8e35d0_Kalman_NBA_Line_up_Analysis.pdf

Riccardi, N. (2023). Optimizing NBA Roster Construction. Academy of Economics and Finance Journal, 14, 28-38. https://surface.syr.edu/cgi/viewcontent.cgi?article=1066&context=sportmanagement