**Analyzing Success Factors for NBA Game ‘Personality’ Clusters**

Basketball

291560

**1. Introduction**

“This is gonna be a slugfest.” - Nick Nurse, ahead of the Raptors 2022 Eastern Conference Playoff’s matchup against the 76’ers [1]. This assertion touches on a conventional wisdom: NBA games have personalities of their own. Some games are high-scoring, edge of your seat affairs while others, as Nurse alludes too, are characterized by hard-fought, physical play. These different games tend to play to the skillset advantages of certain teams and players and contribute to the identities and strategic approaches of those teams. In this paper, we use support vector classification and principal component analysis to visualize the disparate game groups across the features that most contributed to winning. Next, k-means clustering is used to examine advanced box scores from NBA games and identify and assign ‘personality’ types to the games themselves. Finally, we implement Apriori rule analysis to extract the key combinations of success factors that impact winning within each type.

**2. Background**

This paper explores two popular areas of study within basketball research: clustering and success factors. Clustering approaches are generally more interested in archetyping or creating like groups within the larger dataset that indicate divergences in traits such as strategy, skill, or outcome. While not mutually exclusive, those studies interested in determining success factors, or those characteristics which most positively affect winning, typically use some form of discriminant analysis to determine which features from the broader set are the most impactful.

**2.1. Clustering**

Deep learning algorithms, k-means in particular, have been a frequent tool used to derive complex patterns and success factors from NBA data [cite survey]. Primarily, these approaches have used one or both of two primary data sources: player tracking data and box-score statistics.

Clustering player tracking data is generally aimed at identifying the patterns of movement present in coordinated player movements called actions. Brooks [cite] achieves this by building images of each player’s movement on the court within a given possession and then clustering them into 30 groups which were manually reviewed and labeled to indicate the pattern variants. Similar efforts by Nistala [cite] and Guttag and Nistala [cite] seek to analyze these movements on and individual level, by clustering trajectory embeddings of a single player while Stephanos et al. [cite], look at a specific action (the Dribble Hand-off) and implement k-means clustering to identify the sub-variants in execution utilized by different teams and players. Sampaio et al. [cite] use tracking data actions (i.e., pull up shots, elbow touches, catch-and-shoot possessions, etc.) and clustering to identify success factors for all-stars in the NBA.

Those approaches that use box-score inputs as the primary data source are, however, typically more interested in identifying archetypes of players and teams. For example, Lutz et al. [cite] uses common statistics such as field goals and assists to identify ten different player categories. Dehesa et al. [cite] look at regular season and playoff performance across a spectrum of box-score and composite statistics to identify performance profiles for each which were mostly discerned by NET impact (or how well your team performs with you on the floor as opposed to off). Kalman and Bosch [cite] use cluster analysis to examine player tendencies as derived from advanced statistics and shot location to determine a set of 9 player archetypes, and then perform lineup analysis to optimize the potential 5-player combinations. There are also some instances of clustering being performed without box-score or tracking data, for instance, Zhang [cite] uses anthropometric player attributes and playing experience to create player archetypes and associate them with both team and individual success.

While less common, there are a couple examples of clustering being used to discern archetypes within games themselves, by analyzing trends in game statistics. Sampaio et al. [cite], for instance, clusters games based on scoring as a means of identifying team success factors in the Portuguese Professional Basketball League. They then used discriminant analysis to examine winning teams across these clusters, and were able to determine that regular season games were best discriminated by successful free-throws and playoff games by offensive rebounding. More directly, Rocha da Silva [cite] looked at k-means clusters of historical box score data to identify three different eras within NBA regular seasons. They then analyzed the trends in statistical categories and isolated the most discerning between eras, in particular: the emergence and dominance of the three-point shot. Considering the prevalence in literature for identifying archetypes and success factors for players and teams, deep learning clustering techniques have seldom been used to analyze the patterns between games themselves. To the best of our knowledge, this paper presents the first such targeted analysis of game clustering for the expressed purpose of identifying game archetypes.

**2.2. Success Factors**

Some clustering approaches, such as Ibanez et al. and Samapaio [cite], use clustering as a direct means of generating success factors in the Spanish Basketball League and were able to conclude that the best teams were distinguished from the worst primarily through assists, steals, and blocks. This type of clustering, however, is intended to stratify along previously known distinctions such as win/loss record, and is distinct from the unsupervised clustering approaches used in action detection and archetype.

This is a common approach, however, for identifying success factors. Primarily through comparing winning teams to losing teams with the use of discriminant analysis, those primary statistical categories for which success in translates to success on the whole. Even approaches such as Zhang et al. [cite] use this type of analysis to evaluate the performance archetypes generated through deep learning. Samapia, McGarry et al, also use discriminant analysis to identify different performance indicators related to scoring, passing, defense, and all-round game behavior for k-means clusters generated using NBA tracking data.

While not directly related to clustering, Rocha da Silva [cite] summarizes various additional findings related to team and player success factors in basketball, and highlight rebounding, turnovers, free throws, and assists to all be strong indicators of performance. One limitation from these conclusions and similar discussed approaches, however, is the unifactorial nature of their analysis. Basketball is by nature a dynamic, stochastic sport in which one action may have a plethora of downstream consequences. By using Apriori frequent itemset analysis as a means of deriving which success factors in combination have the biggest impact on winning, we, to the best of our knowledge, introduce a methodology for examining the multivariate impact of different in-game statistical factors on team success.

**3. Methods**

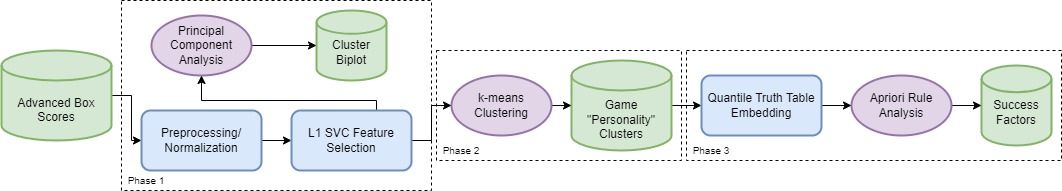
**3.1. Data Overview**

In this paper, we explore advanced box score statistics from the 2014-2018 NBA regular seasons. To avoid redundancy bias, only those entries for home games were considered by the clustering algorithm. In an effort to eliminate noise and reduce the scope of the feature set to be interpreted by the clustering algorithm, L-1 based feature selection using a support vector classifier was implemented with a threshold .02 to reduce the set of features from 35 to 25. A full overview of the statistical features and their role is contained in the table below.

|  |  |  |
| --- | --- | --- |
| Name | Description | Role |
| Team | Three letter abbreviation for the team this box score entry is for, i.e. NYK, DEN | Meta |
| Game | A numeric identifier for the game in question | Meta |
| Date | The date the game occurred, in ‘yyyy-mm-dd' format | Meta |
| Home | Whether or not the game occurred in the Team’s home arena | Meta |
| Opponent | The opponent team, represented by the same three letter abbreviation, i.e. BOS, LAL | Meta |
| WINorLOSS | A ‘W’ or ‘L’ indicating whether the Team won the game | Target |
| TeamPoints | Total amount of points scored in the game by Team | Feature |
| OpponentPoints | Total amount of points scored in the game by Opponent | Feature |
| FieldGoals | Total amount of successful field goals (shots from within the 3pt arc) by Team | Feature |
| FieldGoalsAttempted | Total amount of field goal attempts (including misses) by Team | Ignored |
| FieldGoals% | The percentage of makes to misses in field goal attempts by Team | Ignored |
| 3PointShots | Total amount of successful shots taken from beyond the 3pt arc by Team | Feature |
| 3PointShotsAttempted | Total amount of 3pt shot attempts (including misses) by Team | Feature |
| 3PointShots% | The percentage of makes to misses in 3pt shot attempts by Team | Ignored |
| FreeThrows | Total amount of successful free throws taken by Team | Feature |
| FreeThrowsAttempted | Total amount of free throw attempts (including misses) by Team | Ignored |
| FreeThrows% | The percentage of makes to misses in free throw attempts by Team | Ignored |
| OffRebounds | Total amount of offensive rebounds collected by Team | Feature |
| TotalRebounds | Total amount of rebounds (defensive included) collected by Team | Feature |
| Assists | Total number of assists (passes leading directly to score) by Team | Feature |
| Steals | Total number of steals (taking possession of the ball from the opponent) by Team | Feature |
| Blocks | Total number of blocks (opponent shots physically impacted by defender) by Team | Feature |
| Turnovers | Total number of turnovers (actions that lead to a loss of possession) by Team | Feature |
| TotalFouls | Total number of fouls committed (offensive and defensive) by Team | Feature |
| Opp.FieldGoals | Total amount of successful field goals (shots from within the 3pt arc) by Opponent | Feature |
| Opp.FieldGoalsAttempted | Total amount of field goal attempts (including misses) by Opponent | Ignored |
| Opp.FieldGoals% | The percentage of makes to misses in field goal attempts by Opponent | Ignored |
| Opp.3PointShots | Total amount of successful shots taken from beyond the 3pt arc by Opponent | Feature |
| Opp. 3PointShotsAttempted | Total amount of 3pt shot attempts (including misses) by Opponent | Feature |
| Opp. 3PointShots% | The percentage of makes to misses in 3pt shot attempts by Opponent | Ignored |
| Opp. FreeThrows | Total amount of successful free throws taken by Opponent | Feature |
| Opp. FreeThrowsAttempted | Total amount of free throw attempts (including misses) by Opponent | Ignored |
| Opp. FreeThrows% | The percentage of makes to misses in free throw attempts by Opponent | Ignored |
| Opp. OffRebounds | Total amount of offensive rebounds collected by Opponent | Feature |
| Opp. TotalRebounds | Total amount of rebounds (defensive included) collected by Opponent | Feature |
| Opp. Assists | Total number of assists (passes leading directly to score) by Opponent | Feature |
| Opp. Steals | Total number of steals (taking possession of the ball from the opponent) by Opponent | Ignored |
| Opp. Blocks | Total number of blocks (opponent shots physically impacted by defender) by Opponent | Feature |
| Opp. Turnovers | Total number of turnovers (actions that lead to a loss of possession) by Opponent | Feature |
| Opp. TotalFouls | Total number of fouls committed (offensive and defensive) by Opponent | Feature |

**3.2. Architecture Pipeline**

This research is conducted using a pipeline with 3 primary phases: preprocessing & visualization, k-means clustering, and Apriori association mining, as illustrated in Figure 1.



Normalization and feature selection precede all other steps, while each primary phase has its own distinct deliverable. Previous clustering research in the NBA has focused on play-action [2], player archetypes [3], and historical eras [4]. This is, to the best of our knowledge, the first examination of game personality clustering.

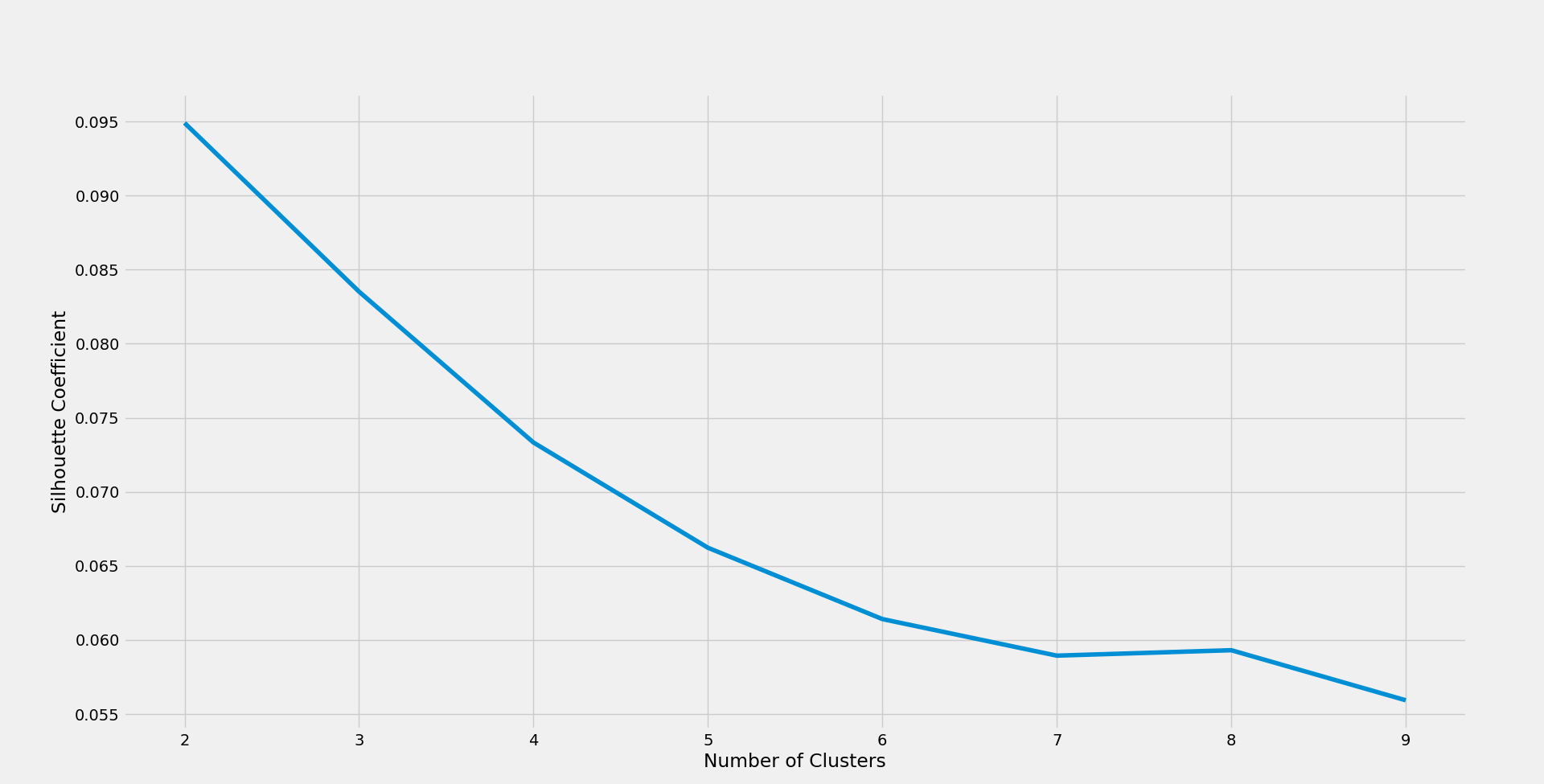
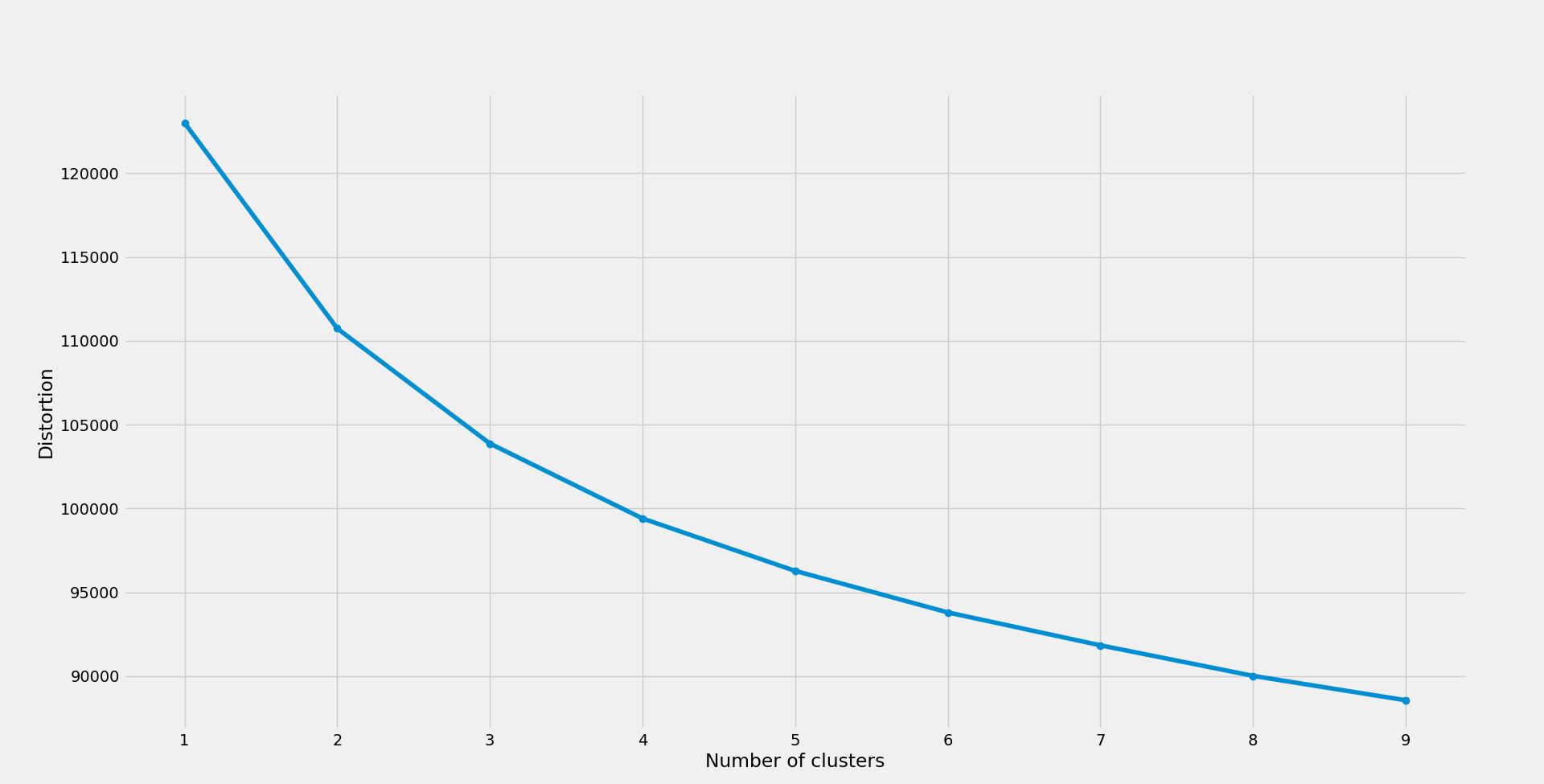
**3.2.1. Principal Component Analysis**

In this paper, we use principal component analysis (PCA) primarily for vis ualizing the clusters of data. PCA is widely used for these purposes, as it seeks to represent the variability of a multi-dimensional dataset in a lower dimensional non-correlated variable set (or principal component) which is a linear combination of the original dataset [cite]. This is particularly useful, as by using the first two orthogonal primary components, we can represent the entirety of the feature set in a two-dimensional graph called a bi-plot. A detailed explanation and the inspiration for this visualization is contained in Rocha da Silva’s work studying historical game clusters. Additional information and an example of the PCA biplot can be found in Section 4.1.

**3.2.2. Clustering**

We implement the scikitlearn k-means clustering algorithm [cite] to group games using in-game statistics collected from advanced box-scores from the 2014-2018 NBA regular seasons [cite]. K-means is a non-hierarchical unsupervised learning technique that seeks to group *n* individuals into *k* pre-determined number of clusters based off their proximity to each other in the multi-dimensional feature space. Johnson and Wichern provide a detailed explanation of this process [cite]. K-means allows for a bottom-up approach to categorization which is ideal for pattern mining.

Data is prepared using pandas dataframes [cite] and normalization is performed on numeric features to avoid scale bias. Elbow plots and silhouette scores were generated to help determine *k.* Both measures iterate over different potential *k* values, and then evaluate the quality of the clusters [cite]. In both cases, a determination of diminishing returns must be deduced, suggesting an ideal *k*-value, which in our case led us to four unique clusters. Both plots can be found below.



Once determined, the model was fit to the data, and the chosen labels were assigned to the box-score entries.

**3.2.3 Association Mining**

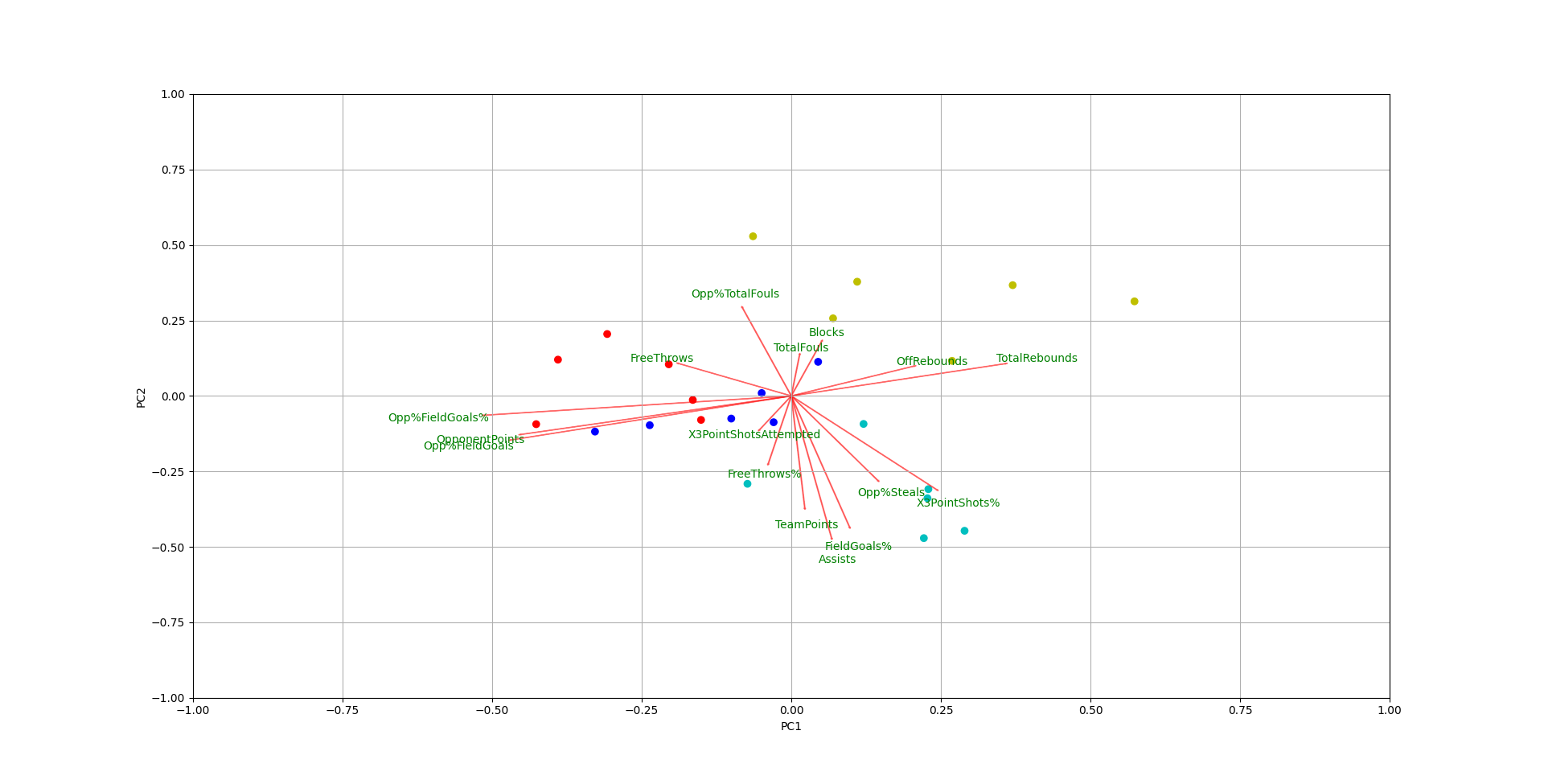
In order to determine those success factors that most impacted winning, we broke each statistical category into quantiles. We then created a truth-table where the rows represent the individual games in the box-score, and the columns contain Boolean values indicating whether the quantile threshold for that statistical category was overcome. The embedded data was passed per cluster to the Apriori algorithm to determine which success factors most impacted winning.

The Apriori algorithm functions by identifying frequently occurring itemset within the samples, and then correlating them to some other associated result. Classically, this type of analysis is referred to as shopping cart analysis, and would allow for conclusions such as: Shoppers who purchased tofu and kale were more likely to also purchase vegan cheese. For our purposes, the target value (or right-hand side) was either winning or losing the game, and the indicators (or left-hand side) were Boolean values indicating presence in the corresponding threshold, such as: 3ptShotAttempts > 80th percentile.

**4. Results**

**4.1. PCA Biplot**

PCA analysis is used to generate the two principal components (PC1, PC2) that represent the x and y axes in the graph below. The various red lines represent the top features as selected by the L1 classifier, and are labeled in green. The various color dots are the samples closest to the center for each cluster and are color coated to match.



As we can see in the graph, certain key indicators occupy similar space within the bi-plot, with interesting groupings such as Assists, FieldGoal%, and TeamPoints as well as Fouls with Blocks and to a lesser extent, 3PointShotsAttempted with FreeThrow%. Additionally, clear distinctions can be found between the clusters, with each roughly occupying one of the four cardinal sections of the graph, thus indicating the relative similarity within and distinction between groups.

**4.2. Clustering**

The four resulting game clusters were then condensed into feature averages and compared to assess the ‘personality’ of each. We then associated the trends in each to one of four colloquial labels from the NBA vernacular: track meet, shooting clinic, blowout, slug fest. Notably, three of the four clusters include a positive scoring margin, lending credence to the theory of home court advantage. Table 1 contains feature summaries for home and away teams of each cluster.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster Label | Score | Field Goals | 3 Point % | Assists | Boards | Fouls | Win % |
| ‘Track Meet’ | 113-112 | 40.3-39.8 | .364-.359 | 23.4-22.5 | 44.6-43.7 | 23.6-24.5 | .561-.439 |
| ‘Shooting Clinic’ | 102-111 | 38.2-42.3 | .342-.428 | 22.1-26.0 | 40.7-42.5 | 17.8-19.5 | .228-.772 |
| ‘Blowout’ | 114-99 | 43.7-37.3 | .433-.326 | 27.9-21.3 | 44.6-40.4 | 18.7-17.9 | .932-.068 |
| ‘Slug Fest’ | 95-93 | 35.4-34.8 | .314-.304 | 20.2-19.1 | 45.8-44.6 | 19.3-20.2 | .596-.404 |

**4.3. Association Mining**

The rules targeting outcome were extracted and sorted by lift (how impactful the rule was within the personality cluster). Our approach is unsupervised, thus unaware that the team with the highest score wins therefore, we manually filter out noisy rules like: Knicks outscored Lakers. These rules provide game insights that we believe, if utilized by a domain expert, would provide a strategic advantage. Table 2 contains a sample of top rules for each cluster.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | Rule | Confidence | Lift | Cluster | Rule | Confidence | Lift |
| 1 | Opp.FieldGoalsAttempts > 93, OppFieldGoals > 43, TotalFouls > 27 | .928 | 1.19 | 3 | Blocks > 7, Opp.FieldGoalsAttemps > 91, TotalFouls > 22 | .936 | 1.21 |
| 1 | FreeThrows > 28, X3PointShots > 32,  TotalFouls > 27 | .877 | 1.12 | 3 | Turnovers > 16, Opp.3PointShots > 11, Opp.FieldGoalsAttempts > 91 | .908 | 1.17 |
| 2 | Opp.FreeThrows > 18, Opp.3PointShots > 15, Opp.FieldGoalsAttempts > 90 | .898 | 1.16 | 4 | Steals > 10,  TotalFouls > 22 | .925 | 1.18 |
| 2 | Turnovers > 16,  TotalFouls > 21 | .914 | 1.18 | 4 | Turnovers > 17, Opp.FieldGoalsAttempts > 89 | .909 | 1.16 |

**5. Conclusion**

This paper presents a novel approach to examining key success factors within different personality trends of NBA games. We believe this technique has cross-domain appeal in other sports, markets, and industries.

**References**

[1] A. Rose, "Raptors Need to Avoid 'Slugfest' by Dictating Tempo to 76ers", *Sports Illustrated Toronto Raptors News, Analysis and More*, 2022. [Online]. Available: <https://www.si.com/nba/raptors/news/toronto-raptors-slugfest-transition-offense-philadelphia-76ers-playoffs>.

[2] D. Stephanos, G. Husari, B. Bennett, M. Harrisson and E. Stephanos, “Using Hex Maps to Classify & Cluster Dribble Hand-Off Variants in the NBA,” in *Proceedings of the MIT Sloan Sports Analytics Conference*, 2022.

[3] S. Kalman and J. Bosch, “NBA Lineup Analysis on Clustered Player Tendencies: A new approach to the positions of basketball & modeling lineup efficiency,” in *Proceedings of the MIT Sloan Sports Analytics Conference*, 2020.

[4] J. Rocha da Silva and P. Rodrigues, "The three Eras of the NBA regular seasons: Historical trend and success factors", Journal of Sports Analytics, vol. 7, no. 4, pp. 263-275, 2021. Available: 10.3233/jsa-200525.

Sampaio, J. , & Janeira, M. 2003, Statistical analyses of basketball team performance: Understanding teams’ wins and losses according to a different index of ball possessions. International Journal of Performance Analysis in Sport, 3(1), 40–49. doi: [10.1080/24748668.2003.11868273](https://dx.doi.org/10.1080/24748668.2003.11868273)

[8] D. K. Stephanos, G. Husari, B. T. Bennett, and E. Stephanos, “Machine Learning Predictive Analytics for Player Movement Prediction in NBA: Applications, Opportunities, and Challenges,” in *Proceedings of the 2021 ACM Southeast Conference*, 2021, pp. 2–8, doi: 10.1145/3409334.3452064.

Kubatko, J. , Oliver, D. , Pelton, K. , & Rosenbaum, D. T. 2007, A Starting Point for Analyzing Basketball Statistics. Journal of Quantitative Analysis in Sports, 3(3), 2007. doi: [10.2202/1559-0410.1070](https://dx.doi.org/10.2202/1559-0410.1070)

[14] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[15] D. Lutz, “A Cluster Analysis of NBA Players,” in *Proceedings of the MIT Sloan Sports Analytics Conference, Boston, MA, USA.*, 2012, vol. 24, p. 2016.

Ibáñez, S. J. , Sampaio, J. , Feu, S. , Lorenzo, A. , Gómez, M. A. , & Ortega, E. 2008, Basketball game-related statistics thatdiscriminate between teams’ season-long success. EuropeanJournal of Sport Science, 8(6), 369–372. doi: 10.1080/17461390802261470

[18] S. Kalman and J. Bosch, “NBA Lineup Analysis on Clustered Player Tendencies: A new approach to the positions of basketball \& modeling lineup efficiency of soft lineup aggregates,” 2020.

[6] J. D. Brooks, “Using Machine Learning to Derive Insights from Sports Location Data,” Massachusetts Institute of Technology, 2018.

[1] A. Nistala and J. Guttag, “Using Deep Learning to Understand Patterns of Player Movement in the NBA,” in *Proceedings of the MIT Sloan Sports Analytics Conference*, 2019, pp. 1–14.

[13] A. Nistala, “Using deep learning to understand patterns of player movement in basketball,” 2018.