Examining NBA Team Chemistry and Success

Group 174: Nicholas Archambault, Armando Di Cicco, Luis Dominguez,
Vinay Easwaran, Alexander Morton, Jere Xu

December 1, 2023

Introduction

This project seeks an answer to the question, 'Does an NBA team's chemistry make a quantifiable impact on its ability to secure a playoff berth?'

It is commonly thought that the success of an organization depends primarily on the talent of the individuals who comprise it. We argue that performance is also influenced by teammates' familiarity with one another, manifested through on-court chemistry that elevates play.

This project's objective is to investigate NBA team chemistry in relation to team performance. An undirected, interactive arc-diagram graph was implemented using Javascript D3 to display teammate connections across various teams and seasons. Novel measures of team chemistry were derived and added to a data set of NBA box score statistics, and classification models predicting whether a team makes the playoffs were built, evaluated, and compared. Models were constructed using the k-nearest neighbors (KNN), support vector machine (SVM), Naive Bayes, and decision tree algorithms for classification, and predictors included the derived chemistry metrics as well as team season statistics.

An undirected arc-diagram graph displaying player relations is a novel tool, in large part because the data necessary for its creation is disjoint, requiring substantial scraping and cleaning to compile. Unifying team chemistry and success in the same visualization permits interactive engagement via pointed queries of specific teams, seasons, and players. This visual tool represents a freshly intuitive demonstration of the impact of a latent but nonetheless powerful factor in team performance. The relationship between team chemistry and team success is captured and quantified in a rigorous manner through a suite of machine learning models predicting a team's playoff membership. Such a framework illuminates deeper user understanding of the sport and offers a springboard for future innovations and insights.

Literature Review

Reviewed below is a subset of existing literature on the psychology of teamwork and the ways machine learning-based classification has been applied to big data design and sports.

Psychology of Teamwork

Teamwork as a socio-psychological mechanism among groups has been examined across several contexts, from IT companies [1] to squads of competitive video game players [2]. Teamwork has been defined as a collaborative effort to maximize a team's likelihood of achieving its purposes [3], and teammate familiarity positively impacts performance in both hierarchical and horizontal group structures [1] [4].

Team chemistry, as a byproduct of healthy communication and collective team experience [3], measurably improves performance outcomes [5]. Strong culture and supportive communication are distinguishable variables that drive team success [6] [7].

Examination of team dynamics in these wider domains provides a solid foundation for their study in sports, an idealized context given the ubiquity of well-defined teams and ease with which success can be measured. A number of studies specifically leverage basketball and its highly cohesive, fluid nature as a test case for establishing Bayesian models for broader applications [8], representing teams as networks of social interaction [9], and defining and measuring teammate synergy and the chemistry-success relationship [10] [5]. The breadth and quantitative nature of such basketball-focused work underscores the virtue of examining team chemistry among NBA players.

Machine Learning Classification in Sports and Big Data Design

The challenges of visualizing large data sets have been well documented [11]. Drawing upon design paradigms [12], frameworks for visualizing complex network data [13] – particularly the use of the fact discovery, fact organization, and visualization modules [14] – work in concert with research from other domains [15] to inform our design decisions with respect to visualizing NBA team chemistry.

Given the importance of managing scalability and computational efficiency when working with large data sets, innovation in data dimensionality reduction is a crucial undertaking [16]. Besides traditional techniques like PCA, the efficacy of new machine learning-focused approaches involving KNN has also been demonstrated [17].

NBA team performance is fertile ground for the leveraging of various statistical and machine learning approaches. Bayesian models have been used to rank individual players and teams [18] [19] and derive age curves to show career-long skill deterioration [8]. The model choices we have made for this project have robust histories of application in sports: KNN and SVM approaches are favorable for grouping players into archetypes [20], while other classifiers have been used to assess team chemistry [21] and predict game outcomes in basketball [22] and other sports [23].

Methodology

This project's exploration of the relationship between NBA team chemistry and success consists of two components. The first is an interactive visualization showing metrics of teammate chemistry. The second is a comparison of the performance of machine learning classifiers, including KNN, SVM, Naive Bayes, and decision trees (CART), in using team chemistry metrics alongside traditional team season statistics to predict whether a team earns a playoff berth.

This project builds on a substantial body of work analyzing teamwork and success within the dynamic, cohesive context of basketball, but it puts forth a number of key innovations in its approach. There is no existing tool that presents a graph representation of how NBA teammates are connected across nearly two decades, nor is there a resource bringing together metrics measuring both teammate familiarity and overall team success. The derived chemistry metrics – which include cumulative teammate minutes, games, and seasons played together – are straightforward but novel representations of team chemistry, requiring a process of data compilation from disparate sources that has not been previously implemented. Furthermore, this project uses machine learning classifiers to examine how the chemistry-success relationship translates to playoff membership. The performances of KNN, SVM, Naive Bayes, and CART frameworks were assessed on baseline models which excluded the derived team chemistry metrics as well as full models that included them. For every classifier, the full model outperformed the baseline, suggesting that novel metrics quantifying team chemistry can contribute to more accurate projection of team success than can

traditional statistics alone. This finding validates the hypothesis that successful team performance is dependent not only on game and statistical outcomes, but also historical teammate experience and on-court familiarity.

Data

Data were gathered primarily from existing Kaggle sources. Three data sets – relating to game listings and game details from the 2004 season onward, as well as player career and biographical details – were downloaded from Kaggle. One additional data set detailing team season outcomes was scraped from the data website 'Land of Basketball.'

These data sets were merged to use in the visualization tool, then leveraged in the machine learning models. The data contained derived team chemistry metrics, including total teammate minutes, games, and prior seasons played together for every NBA player across every season from 2004 onward. This temporal-ordered data set was merged with downloaded Kaggle data that compiled team season statistics and playoff outcomes. The final data set used for model building and analysis had dimensions of 575 rows and 26 columns, occupying 60 MB on disk. It consisted of team season statistics, the binary variable indicating succesful playoff berth, and the engineered chemistry metrics for each team and season from 2004 to 2022.

Visualization

After a substantial process of cleaning and manipulating input data, the interactive arc-diagram graph visualization was created using Javascript D3. The web page hosting this tool contains two drop-down menus, allowing a user to select their desired team and season for examination. Under the drop-downs, the team's record, regular season power ranking, and playoff outcome are shown alongside its team chemistry metrics: how many minutes and games its players previously played together, and how many total previous connections (arcs) exist among its players.

When the cursor is not hovered over a specific graph node, all nodes and arcs are visible. Each arc denotes a player's connection to a teammate, indicating they had played together at some point in their careers prior to the selected season. When a specific player node is moused over, that player's arcs are isolated and all others are hidden. On the right-hand side of the screen, the player's name is displayed above the names of teammates with whom he had previously played. For each of those teammates, the number of seasons, games and total cumulative minutes played together are shown.

At the top of the page, a mouse-over element provides disclaimers regarding data sources, explanations of data formatting choices, caveats pertaining to the 2020 season impacted by COVID, and definitions for how team chemistry metrics were calculated.

The tool presents innovative measures of team chemistry in a novel format that allows users to explore insights partitioned not only by team and season, but also by individual player. Thanks to the tool's dynamic nature, users can easily toggle between pages to visually understand questions that previously would have required substantial data gathering, manipulation, and formatting to address. For example, users could examine the total number of minutes played together for NBA championship winners relative to last-place league finishers. Users could isolate a single team to examine how rosters and team chemistry metrics fluctuate in accordance with team success through the years. Users could even track a single player, such as LeBron James, an outstanding player who is known for attracting high-caliber, familiar talent to his teams, and examine how his most familiar teammates have come and gone from his side across various squads and seasons.

The visualization tool is a navigable laboratory for examining how rosters change over time and

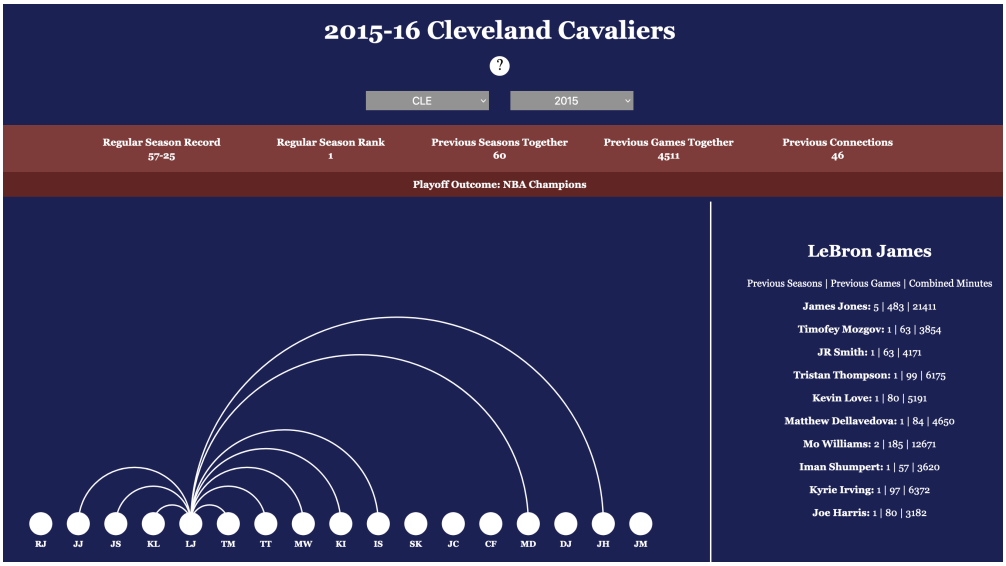

Figure 1: Screenshot of the visualization tool highlighting LeBron James and the 2015-2016 Cleveland Cavaliers, who won the NBA championship.

how team chemistry measures correspond to team success, providing a foundational understanding of the chemistry-success relationship and giving users a starting point to uncover new insights.

Machine Learning Modeling

The algorithmic component of this project used four different machine learning classifiers – KNN, SVM, Naive Bayes, and CART – to better understand the chemistry-success relationship. Two sets of models were constructed, each predicting the binary response outcome of whether a team made the playoffs. Predictors for the 'base' set of models included 22 traditional team season statistics, such as points per game, field goal percentage, free throws attempted per game, and other common box score stats. Predictors for the 'full' set of models included those 22 team statistics plus four derived chemistry metrics: total minutes, games, and seasons played together by all teammates prior to the season at hand, and total games played together by all teammates during that given season. These metrics offer an innovative addition to typical assessments of team performance, which tend to rely exclusively on granular in-game statistics, in order to directly evaluate the research question of whether team chemistry elevates team success.

The two sets of models were created in order to isolate the effects of the chemistry metrics relative to the predictive power of in-game statistics alone. The KNN, SVM, Naive Bayes, and CART algorithms were selected in order to explore parametric versus non-parametric approaches as well as linear versus nonlinear decision boundaries.

Experimentation and Results

Experimentation in the algorithmic component of this project sought to understand whether certain classification algorithms were better than others at predicting team success; whether a set of model predictors that included chemistry metrics could more accurately predict team success across multiple model types than a data set solely consisting of box score statistics; and exactly how much predictive power such chemistry metrics held in comparison to box score statistics. A combination of box score statistics and derived chemistry metrics were used to predict whether a team made

Algorithm	**Hyperparameters**	**Train Acc. (base data)**	**Test Acc. (base data)**	**Train Acc. (full data)**	**Test Acc. (full data)**
KNN	*neighbors* = 31 *weights* = uniform	0.746	0.765	0.782	0.8
SVM	$C = 5$ *kernel* = linear	0.848	0.861	0.872	0.887
Naive Bayes	n/a	0.67	0.67	0.759	0.765
Decision Tree	*max depth* = 8 *min. samples leaf* = 16 *max features* = 11 *min. samples split* = 35	0.778	0.678	0.807	0.704

Table 1: Table showing results of testing across model types and data sets.

the playoffs and thus achieved a successful season. Although standards vary between stronger and weaker teams, and also among various team stakeholders – the players' definition of 'success' might be different than that of the front office or fan base – making the playoffs is generally considered a successful NBA season outcome and was interpreted as such for this project.

After preprocessing and merging, the data were split into train and test sets with an 80% split, then normalized to ensure proper conditions for the KNN and SVM algorithms, which are sensitive to scale. The KNN, SVM, Naive Bayes, and CART models were first run on the base data set consisting only of traditional NBA box score stats. For each model a subset of hyperparameters was tuned using grid search. Each model was subjected to 10-fold cross validation in order to mitigate overfitting. The model whose hyperparameter combinations yielded the best train accuracy was fitted to the train data, then used to predict on the test data.

This process was then repeated using the full data set, which included as four additional predictors the derived team chemistry metrics. Using the modified data set, the hyperparameters were re-tuned, and cross validation, fitting, and prediction were conducted again.

Reasonable train and test accuracies were achieved in this initial testing, but the process was prone to severe overfitting, so analysis was reworked. In the second phase, optimal hyperparameters found using the full data set were also used to build models for the base data set; hyperparameters were not re-tuned separately for each data set. This methodology served as an attempt to protect against overfitting and allowed for more direct comparison between base and full results. Additionally, the decision trees – the model type most overfitted in the first round of experimentation – were rerun using a broadened grid search over a wider set hyperparameters.

These experimental changes led to substantial improvement in model fits relative to initial testing. Results summarizing the modified phase of experimentation are shown in Table 1.

The overall top-performing model iteration was the SVM, which yielded a test accuracy of 0.887 on the full data. All model types experienced a moderate to substantial increase in both train and test accuracy when the full data set was used instead of the base, supporting the hypothesis that the inclusion of derived team chemistry metrics leads to better projection of playoff outcomes. Slight overfitting was still evident, but experimentation throughout the modeling process substantially reduced overfitting and led to sounder results relative to initial testing.

The decision trees were useful in determining the predictive power of the derived chemistry metrics relative to other box score predictors. Table 2 shows the top six feature importances for the base and full decision trees.

The feature importance values revealed the predictive power of the chemistry metrics and

Base Data	*Full Data*
field goal pct.: 0.415	**games together: 0.362**
turnovers: 0.184	field goal pct.: 0.311
3-point pct.: 0.13	defensive rebounds: 0.058
field goals attempted: 0.091	**total minutes: 0.057**
free throws made: 0.07	3-point field goals made: 0.035
points: 0.04	free throws attempted: 0.034

Table 2: Feature importance of decision tree models. Derived chemistry metrics are in bold.

underscored the notion that measures of team chemistry effectively predict team success. For the decision tree run on the full data set, two of the top six most important features – including the top one – were chemistry metrics. The top feature, games played together among teammates, had an importance value greater than that of field goal percentage, which is considered the cornerstone of effective NBA offense. Taken along with evidence that model performance generally improves relative to a baseline when chemistry metrics are introduced, these results support this project's hypothesis and suggest that team chemistry and success are positively intertwined.

Conclusions and Discussion

This project affirmatively answers the question of whether derived measures of NBA team chemistry can more effectively predict team success than on-court statistics alone. An interactive arc-diagram visual permits users to develop a foundational understanding of the chemistry-success relationship by viewing the cumulative minutes, games, and seasons played together by teammates. This tool assembles a data set collected from disparate sources, calculates novel proxies of team chemistry, and presents information on chemistry and team success in an original interface that opens the door to new insights.

Four classification algorithms were run on a baseline data set without chemistry metrics and subsequently on a full data set that included them. The process revealed that model accuracy improves when team chemistry is considered, and that chemistry metrics play an important predictive role relative to other features. These results indicate that team chemistry makes a quantifiable impact on a team's chances of earning a playoff berth. More broadly, they provide a foundation for future work on how the subtle importance of team chemistry can be measured as a driver of season-long success. Extensions to this work could improve on modeling practices by constructing more powerful model frameworks, like random forests and neural networks, or deriving more rigorous measurements of team chemistry to incorporate into such models.

References

[1] Robert S. Huckman et al. Team familiarity, role experience, and performance: Evidence from indian software services. *Management Science*, 55(1):85–100, 2009.

[2] Colin DeLong et al. Teamskill: Modeling team chemistry in online multi-player games. In *Advances in Knowledge Discovery and Data Mining*, page 519–531, 2011.

[3] Desmond McEwan and Mark R. Beauchamp. Teamwork in sport: a theoretical and interrogative rewiew. *Journal*, 7(1):229–250, 2014.

[4] Gangmin Son, Jinhyuk Yun, and Hawoong Jeong. Quantifying team chemistry in scientific collaboration. *arXiv preprint arXiv:2202.07252*, 2022.

[5] Albert V Carron et al. Team cohesion and team success in sport. *Journal of sports sciences*, 20(2):119–126, 2002.

[6] Benjamin Salcinovic et al. Factors influencing team performance: What can support teams in high-performance sport learn from other industries? a systematic scoping review. *Sports Medicine - Open*, 8(1), 2022.

[7] André Lachance and Jean François Ménard. *Team Chemistry: 30 Elements for Coaches to Foster Cohesion, Strengthen Communication Skills, and Create a Healthy Sport Culture*. ECW Press, Toronto, Ontario, Canada, 2022.

[8] N. Vaci, D. Cocić, B. Gula, et al. Large data and bayesian modeling—aging curves of nba players. *Behav Res*, 51:1544–1564, 2019.

[9] André Lachance and Jean François Ménard. *Team Chemistry: 30 Elements for Coaches to Foster Cohesion, Strengthen Communication Skills, and Create a Healthy Sport Culture*. ECW Press, Toronto, Ontario, Canada, 2022.

[10] Allan Maymin et al. Nba chemistry: Positive and negative synergies in basketball. *SSRN*, 2011.

[11] Makrufa Hajirahimova and Marziya Ismayilova. Big data visualization: Existing approaches and problems. *Problems of Information Technology*, 09:72–83, 2018.

[12] Paul Parsons. Understanding data visualization design practice. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):665–675, 2022.

[13] H. C. Purchase et al. An experimental study of the basis for graph drawing algorithms. *ACM Journal of Experimental Algorithmics*, 2:4, 1997.

[14] Qing Chen et al. Calliope-net: Automatic generation of graph data facts via annotated node-link diagrams. *arXiv preprint arXiv:2308.06441*, 2023.

[15] Lisa Perkhofer et al. Does design matter when visualizing big data? an empirical study to investigate the effect of visualization type and interaction use. *Journal of Management Control*, 31(1-2):55–95, 2020.

[16] Nikos Bikakis and Timos Sellis. Exploration and visualization in the web of big linked data: A survey of the state of the art. *arXiv preprint arXiv:1601.08059*, 2016.

[17] Jian Tang et al. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th international conference on world wide web*, 2016.

[18] Jian Shi and Xin-Yu Tian. Learning to rank sports teams on a graph. *Applied Sciences*, 10(17):5833, 2020.

[19] Edgar Santos-Fernandez, Paul Wu, and Kerrie L. Mengersen. Bayesian statistics meets sports: a comprehensive review. *Journal of Quantitative Analysis in Sports*, 15(4):289–312, 2019.

[20] M. Beckler, H. Wang, and M. Papamichael. NBA Oracle. 2008–2009, 2013. Zuletzt besuchtam.

[21] Prastuti Singh and Bai Yang Wang. NBA Game Predictions based on Player Chemistry, 2019.

[22] Rory Bunker and Teo Susnjak. The application of machine learning techniques for predicting match results in team sport: A review. *Journal of Artificial Intelligence Research*, 73:1285–1322, 2022.

[23] Engin Esme and Mustafa Servet Kiran. Prediction of football match outcomes based on bookmaker odds by using k-nearest neighbor algorithm. *International Journal of Machine Learning and Computing*, 8(1):26–32, 2018.