

NBA Team Chemistry

Interactive Arc Diagrams & Machine Learning Algorithms

Luis Dominguez Vinay Easwaran Nicholas Archambault Armando Di Cicco Alex Morton Jere Xu

Introduction

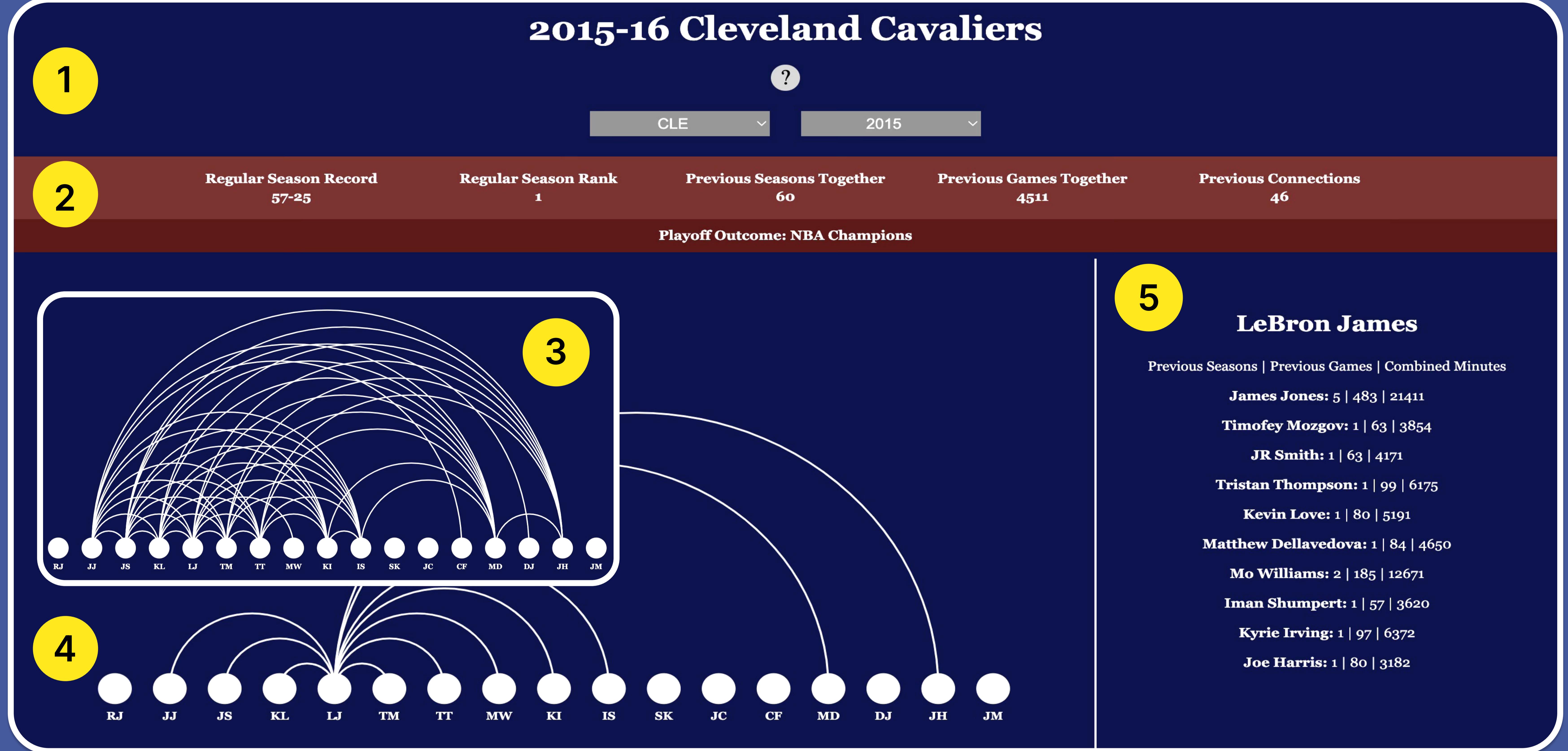
Our investigation into NBA team chemistry and its impact on season performance has two components: an undirected, interactive arc diagram for users to explore chemistry metrics between players, and machine learning models to quantify the impact a team's chemistry has on making the playoffs. It is commonly thought that the success of a team relies on the talent of select individuals. We sought to explore whether overall team member familiarity (chemistry) positively impacts performance, an exploration that has merit in any team-based context, even beyond sports.

Data & Visualization

We downloaded datasets from Kaggle containing temporal NBA game data, and scraped data from other historical NBA archives to obtain team statistics and roster information. Most of these input datasets exceeded 5MB in size, with the largest containing about 650,000 rows. We used Pandas and R to clean and create pairwise metrics within the data to create the visualization and machine learning components.

Our visualization uses JavaScript D3 to create an interactive, undirected arc diagram graph that displays links between teammates with previous experience playing together. Hovering over a node singles out a player's connections and displays additional metrics using D3 tooltips. The dashboard containing team-based metrics uses pure HTML and performs dynamic calculations using JavaScript to display updated metrics based on drop-down selections. This tool merges team chemistry and team success into a single visualization, allowing interactive exploration through targeted queries of specific teams, seasons, and players. It provides a freshly intuitive portrayal of the tangible impact of team chemistry that is not easily quantifiable or typically visualized.

2015-16 Cleveland Cavaliers



Above is the user interface displaying a metric dashboard and interactive arc diagram for the 2015-2016 Cleveland Cavaliers. The user selects a team and season from the dropdowns to update the metrics and full arc diagram. The user can then hover over a node, in this case LeBron James (LJ), to display pairwise player chemistry metrics and a reduced arc diagram.

(1) Dynamic Header: The top panel contains dropdowns for Team and Season, updates the title accordingly, and contains a help menu that explains how chemistry metrics are defined. The default display when launching the application is the 2022-2023 Atlanta Hawks. **(2) Dynamic Dashboard:** The panel contains aggregated team chemistry metrics and playoff result. **(3) Full Graph Before Selecting Node:** The default visualization panel contains a fully linked arc diagram between player nodes, indicating which players were on a team together prior to the selected season. **(4) Reduced Graph After Selecting Node:** An updated visualization panel is displayed when a node is hovered over, reducing the linkages to only those associated with the selected node and triggering the information pop-up on the far right. **(5) Information Pop-Up After Selecting Node:** The pop-up Information panel populates with individual player connections and associated chemistry metrics.

Algorithms & Results

This project used KNN, SVM, Naive Bayes, and decision tree models to better understand the chemistry-success relationship. Two sets of models were constructed, one which included only common box score statistics and the other which included these stats as well as derived chemistry metrics. Each model paradigm predicted the binary response outcome of whether a team made the playoffs.

The data were split into train and test sets with an 80% split, then normalized. 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 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 inclusion of derived team chemistry metrics leads to better projection of playoff outcomes.

The feature importance scores revealed the predictive power of the chemistry metrics and further 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, suggesting team chemistry metrics are valuable predictors of team success -- perhaps more so than many box score stats.

Experiments

Experimentation 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 than a data set solely consisting of box score statistics; and how much predictive power such chemistry metrics held in comparison to box score statistics.

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

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

Initial testing led to severe overfitting, so analysis was reworked. In modified testing, hyperparameters were not separately re-tuned for each data set. This served as an attempt to protect against overfitting and allow for more direct comparison between base and full results. Additionally, the decision tree models -- the model type most overfitted in the first round of experimentation -- were rerun using a broadened grid search over more hyperparameters. These experimental changes led to substantial improvement in model fits, revealing that model accuracy improves when team chemistry is considered, and that chemistry metrics play an important predictive role relative to other features.