Significant Factors for Winning NBA Games

Joshua Gross, Yufan He, Anthony Vo
STATS 295

Background

The success of a basketball game depends on a multitude of factors, including individual performance, team coordination, and tactical execution. Traditional analysis methods are often based on basic metrics such as field goal attempts (FGA), assists (AST), rebounds (OREB, DREB), and fouls (PF). While these metrics provide a straightforward description of a team's performance, they often fall short of capturing the deeper dynamics of teamwork and strategy execution due to the complexity and dynamic nature of the game.

To address these limitations, advanced data analysis has increasingly incorporated additional variables in recent years, such as turnovers (TOV), steals (STL), and free throw attempts (FTA). These metrics offer a more comprehensive quantification of key dynamics in the game, allowing for a more detailed understanding of offensive and defensive efficiency, decision-making quality, and the nuances of strategic execution.

This study focuses on key variables, emphasizing shooting decisions (FG2A, FG3A, FTA), rebounding ability (OREB, DREB), team coordination (AST, TOV), and defensive mechanisms (PF, STL). The intricate interactions between these variables reveal the synergy among offense, defense, and strategic execution, as well as their profound impact on game outcomes.

The goal of this study is to systematically analyze these variables to explore their relative importance in determining team success and uncover hidden patterns within the data. This approach not only provides a scientific foundation for the precise evaluation of team performance but also serves as a valuable reference for optimizing tactical designs and enhancing team competitiveness.

Motivation

In the modern field of sports betting, studying the key factors influencing team success holds both academic and practical value. Accurately identifying these factors not only provides bettors with more precise predictive capabilities against sportsbooks but also helps sportsbooks optimize their data models to enhance profitability. At the same time, for teams, improving their winning rates not only strengthens their competitiveness but also effectively expands their fan base and increases commercial value. Therefore, analyzing and uncovering the core factors that determine team success is not only an important direction for academic research but also a key practice for driving the development of the sports industry.

The rationale behind using causal inference for this question is that it is difficult to map the relationships between strategic decisions quantitatively using other methods. While classifiers can help determine whether a team wins or loses, they cannot reveal the full picture of what goes on in a match. For example, a team who wants to emphasize offensive rebounds might want to know what strategic decisions can positively impact the number of offensive rebounds.

Methodology

The data we gathered was from the NBA during and after the so-called Three Point Revolution beginning in 2015. The reasoning behind using this time period was because strategies shifted greatly; teams focused on shooting more three pointers than ever before, meaning that the tactics they used evolved accordingly. As such, games from the 2015-2016 to the 2023-2024 season were used in our analysis.

In the data processing phase, we first performed data cleaning by removing columns unrelated to the analysis objectives. These included composite metrics such as minutes played (`MIN`), total points scored (`PTS`), total rebounds (`REB`), as well as efficiency-related variables (`FG_PCT`, `FG3_PCT`, `FT_PCT`), since these are derived from other variables and carry redundant information. We retained independent variables, such as field goal attempts, rebounds, assists, and turnovers, to facilitate deeper analysis.

In the feature engineering step, we created a new variable, **`FG2A`** (two-point field goal attempts), calculated using the formula:

FG2A=FGA−FG3A

This variable provides insights into a team's shooting choices and strategic preferences, offering a more detailed understanding of game dynamics. However, after adding `FG2A`, we also had to remove `FGA` to get rid of collinear features. In addition to `FG2A`, we added `HOME` as a variable to see whether or not a team being at home had effects on other features.

In our analysis, we used both causal inference and classification in order to determine the most important factors for winning games.

First, we used Pandas and Seaborn tools to compute the correlation matrix between variables and visualized it through a heatmap. This approach allowed us to quickly identify correlations between variables, providing a reference for subsequent modeling.

Next, for causal inference modeling, we utilized the No-Tears algorithm via the CausalNex Python library to construct a Directed Acyclic Graph (DAG). This process was based on standardized numerical data, automatically generating all possible causal relationships (directed edges). To ensure the reliability of the model, we applied a weight filtering criterion, retaining only edges with an absolute weight greater than 0.5, resulting in a significant causal relationship network. This model quantified the direct or indirect influence between variables, providing data support for optimizing team strategies.

To make the results more intuitive, we used PyVis to create an interactive causal network graph. In this graph, each node represents a variable, and each edge indicates a causal relationship, with its weight reflecting the strength of causality. The generated dynamic HTML file was saved as `nba_network.html`, allowing users to interactively view the causal network through a browser. This series of methods provides strong support for understanding the relationships between variables and their impact on game outcomes.

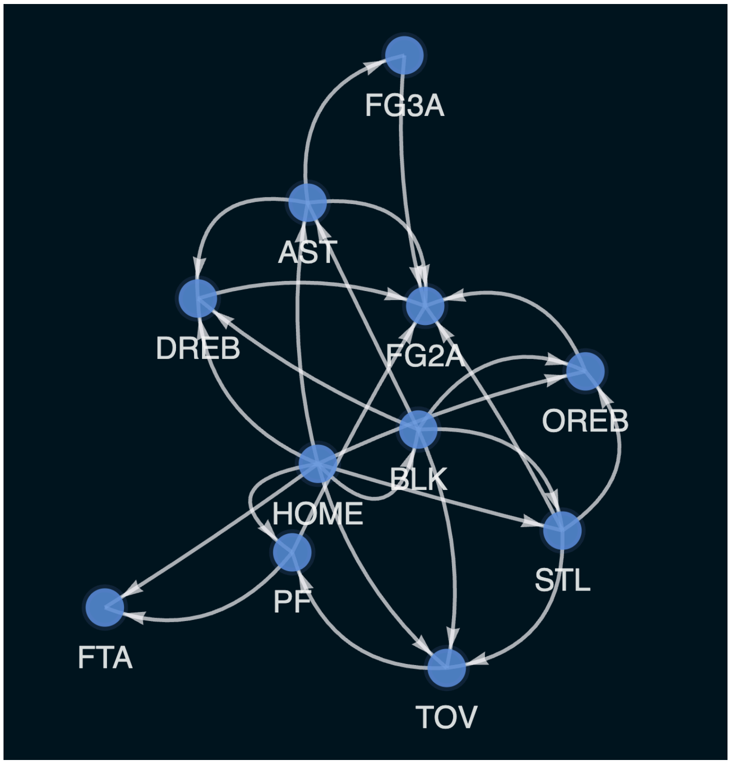
In order to find the variables that have the most impact on winning, we ran a decision tree classifier on our data in addition to running causal inference. We began by scaling our data, then splitting it into training and testing sets. The reason why we ran a decision tree classifier is because it is incorrect to put outcome variables into causal inference. Theoretically, all of the variables we used inside to determine causal inference would have directed edges towards the outcome variable (winning or losing, in this case), so there's no point of including the outcome variable. Therefore, using a classifier makes the most sense to find relationships between the outcome variables and our features.

Results

In the causal network analysis, the No-Tears algorithm helped identify complex causal relationships among variables. FG2A had a significantly high in-degree; AST, FG3A, OREB, DREB and STL were all identified as causing FG2A. Because two-point attempts are typically very efficient, it is important to prioritize them. Thus, strategies prioritizing the causes of FG2A will be more optimal.

HOME was a node with a high out-degree in the causal graph. It had directed edges to DREB, AST, OREB, STL, TOV, PF, and FTA. This implies that whether a team is at home or away heavily affects their counting statistics.

There are many other, singular relationships that show interesting results. FG3A causing FG2A supports the conclusion that shooting three-point shots allows teams to get more two-point shots due to the concept of "spacing" where defenders must leave closer shots open because of the threats posed by three-point shooters. PF having an edge to FTA reveals a correlation between the two, highlighting the fact that referees can have an impact on games by the number of fouls they call. TOV causing PF proves the mental side of basketball: teams that turn the ball over, an action that typically happens through making mistakes, will foul more, an action that typically happens out of frustration or anger.



Causal graph of gathered variables.

In the decision tree constructed from the data, the root node was defensive rebounds (DREB), which highlights their importance in predicting game results. Specifically, when DREB > 0.31, the team demonstrated strong defensive performance, laying the foundation for victory, whereas when DREB ≤ 0.31, the team's defensive control was weaker, significantly reducing the probability of winning.

In both branches, assists and steals emerged as key variables. If assists and steals were low, the team lacked offensive efficiency and defensive transition capability, resulting in a high probability of failure. Even with improved assists, reducing turnovers or enhancing defensive performance was necessary to increase the chance of winning.

In the higher-level branches, defensive rebounds and assists underscored the synergy between defensive and offensive efficiency. In the lower-level branches, steals, turnovers, and personal fouls further illustrated how execution details impacted game results. The leaf nodes of the decision tree clarified how specific combinations of variable thresholds could significantly increase or decrease the likelihood of winning, providing clear and interpretable rules for strategic decision-making.

In the feature importance analysis, we quantified the contribution of each variable to predicting game outcomes based on the decision tree model. The analysis revealed that defensive rebounds ranked first, highlighting their importance in limiting opponents' scoring opportunities and creating counterattacking chances. Assists followed closely, further validating the critical role of team coordination in determining game results. Three-point, two-point, and free throw attempts had moderate importance, indicating the significance of generating opportunities to score. In contrast, whether a team was at home or not had the least importance. This means that a team's ability to win is not influenced much by the location they play in as opposed to the strategies they employ.

In summary, these analyses further deepened our understanding of the factors and mechanisms influencing game outcomes, building upon the insights from the causal network. By integrating variable importance rankings, hierarchical decision paths, and causal relationships, we established a comprehensive analytical framework. This approach not only identifies the key factors affecting wins and losses but also provides actionable and practical recommendations for strategic optimization, offering robust support for teams in tactical planning and performance improvement.

Future Work

Though this project revealed vital information about what the optimal strategies of basketball teams are, there are various improvements that can be made.

One improvement that can help the prediction of these models is the use of advanced statistics. This project only used basic counting statistics, but there are a host of advanced statistics that can glean more information about optimal strategies. A simple example is using a more advanced shot chart; instead of just using FG2A and FG3A, we can discriminate between different FG3A (such as corner vs. above-the-break) or different FG2A (such as mid-range vs. at-the-rim). Thus, more advanced statistics and different datasets can produce more optimal strategies than we can produce with this data.

Another improvement would be to use a larger number of features. Currently, we use team-wide counting statistics as opposed to player-specific ones. It would be more accurate to break up these counting statistics by position and determine which team would win depending on how individual players match up. However, this comes at the cost of a very large dataset: each match would consist of ten players and around ten statistics for each player, leading to a total of one hundred features per match. It is unfeasible to run an accurate and timely causal analysis with this many features with current technologies, but future papers could reveal better methods for causal analysis.

Finally, more sophisticated causal analysis could have been performed. Our model and thought process was flawed, as finding a causal graph between variables does not translate to finding the most important factors to winning. A better model using the methods learned in this class would be a causal policy learner that received rewards for correctly predicting the outcome of a basketball match. With this, we would not need to train a decision tree classifier and could have generated more robust results. We could have also related this policy learner more to the initial motivation of sports betting. A policy learner who maximizes the amount of expected profit per basketball match by betting on money lines would be theoretically analogous to a perfect bettor, revealing how one could derive edges against sports books.

Conclusions

While flawed, our project revealed causal relationships between basketball statistics and highlighted the most relevant features towards winning a match. We found that an optimal basketball strategy would emphasize defensive rebounding and assists, two actions that involve coordination on both the defensive and offensive ends, and that these two actions lead to more high-efficiency shot attempts. Our causal graph also showed how the advent of increased three-point attempts allowed teams to get closer to the basket and take more two-point attempts, a symptom of spacing after the Three Point Revolution.

Overall, we greatly enjoyed working on this project and taking STATS 295 this quarter. Thank you for reading this report, and have a great winter break!