NBA Game Win Prediction Using Multivariable Regression Analysis

2025-11-06

NBA Game Win Prediction Using Multivariable Regression Analysis

Abstract

This analysis investigates the most significant statistical predictors of NBA game outcomes using team-level data from the 2022-2023 season. The core objective is to determine how differential performance metrics between home and away teams influence game results. Our models utilize five key differential variables as predictors in a dual-modeling framework. These include: field goal percentage (FG_diff), free throw percentage (FT_diff), three-point percentage (FG3_diff), assists (AST_diff), and rebounds (REB_diff).

A multivariable logistic regression model is used to predict the binary outcome of a home team win, yielding the probability of victory. Preliminary findings indicate that these performance differentials offer strong predictive power. Specifically, differences in shooting efficiency (Field Goal and Three-Point Percentage) emerge as critical statistical variables influencing game success.

Introduction

The purpose of this analysis is to develop predictive models that can identify which game statistics are most influential in determining the outcome of NBA games. The central research question we aim to answer is: **Which game statistics are the most significant predictors of winning an NBA game?**

This analysis may benefit several groups:

- **Sports analysts and statisticians** seeking to understand the quantitative factors that drive game outcomes
- **Team management and coaches** who can use these insights to identify areas for strategic focus and improvement
- **Sports betting analysts** interested in data-driven game predictions
- **Researchers** studying the statistical relationships in professional basketball

The goal of this project is to train predictive models using game statistics. Specifically, we aim to:

1. Build a logistic regression model to predict the binary outcome of whether the home team wins or loses
2. Identify which statistical differentials (field goal percentage, free throw percentage, three-point percentage, assists, and rebounds) are most significant in predicting game outcomes

By analyzing differential statistics rather than absolute values, we can directly compare team performance within each game and identify which performance gaps are most predictive of victory.

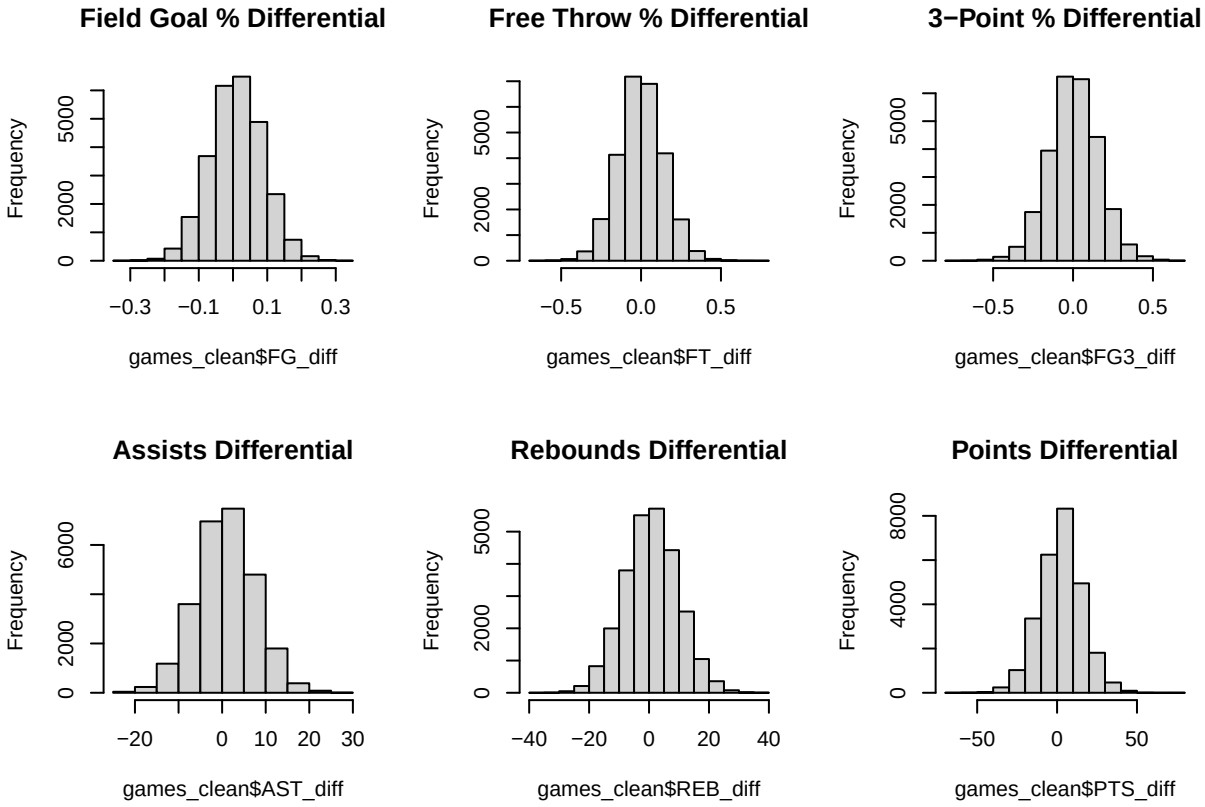
Data

The data used in this analysis was sourced from Kaggle, and comes from two primary datasets. The first dataset “games.csv”, includes game-level statistics for NBA matchups with performance metrics for both home and away teams. These included field goal percentage (FG_diff), free throw percentage (FT_diff), three-point percentage (FG3_diff), assists (AST_diff), and rebounds (REB_diff) per team from the 2022-2023 season. The second dataset “teams.csv”, provides team identification details such as team IDs, abbreviations, nicknames, and cities.

Data Cleaning Steps

The data cleaning process was created through several key steps. First, the team information from "teams.csv" was integrated into the games' data, adding abbreviations, nicknames, and cities for both home and away teams, resulting the variables: "HOME_ABBREVIATION", "VISITOR_ABBREVIATION", "HOME_NICKNAME", "VISTOR_NICKNAME", "HOME_CITY" and "VISITOR_CITY". Next, differential variables were calculated to enable the direct comparison of team performance within each game, including "FG_diff", "FT_diff", "FG3_diff", "AST_diff", "REB_diff", and "PTS_diff", each representing the difference between home and away statistics. During data cleaning, the "GAME_STATUS_TEXT" column was removed as it was unnecessary for analysis, and a new variable, "ABBREVIATION_TEAM_WIN", was created to indicate the winning team by abbreviation. Together, these steps produced a comprehensive data set called `games_clean` containing all original game observations enriched with team details and differential performance metrics, preparing it for further modeling and analysis. Two separate data sets were created as a subset of `games_clean`: `test_data` and `train_data`. `test_data` is comprised of 70% of `games_clean` while `train_data` is comprised of 30% of `games_clean`.

Visualization

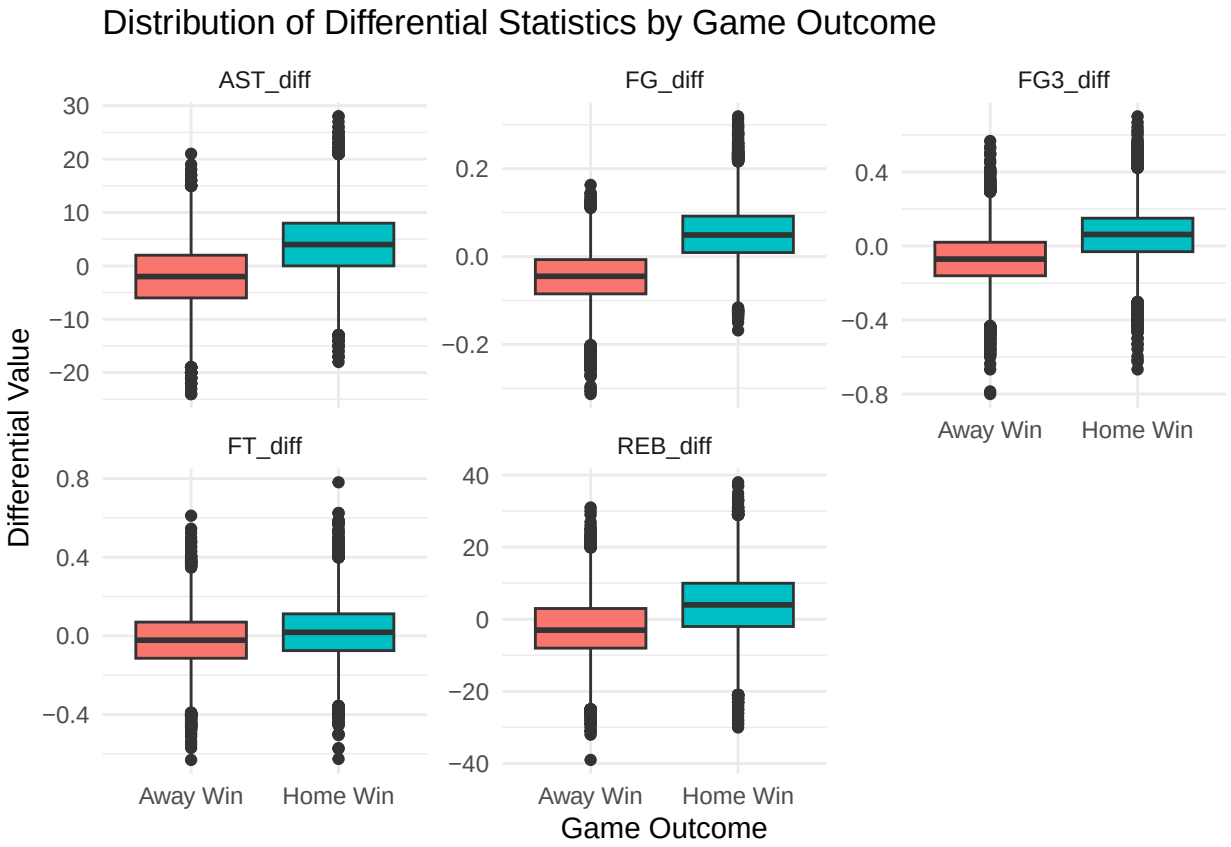


The histograms above display the distributions of each performance differential variable used in the analysis. All six differentials are centered close to zero, which is expected because they represent the home team's performance minus the away team's performance. Each variable shows a roughly symmetric, bell-shaped distribution, indicating that home teams do not consistently outperform away teams in any single metric across all games.

Field goal (FG%), free throw (FT%), and three-point percentage (3PT%) differentials exhibit relatively tight spreads, reflecting that shooting efficiency differences between teams tend to be modest. In contrast, assists and rebounds display wider variability, suggesting larger game-to-game fluctuations in these counting statistics. Point differential has the broadest spread, which aligns with variation in final game outcomes.

These histograms confirm that the predictors behave like continuous, well-distributed performance metrics

without extreme skewness or heavy outliers. This supports the suitability of using these variables in a logistic regression framework to predict game outcomes.

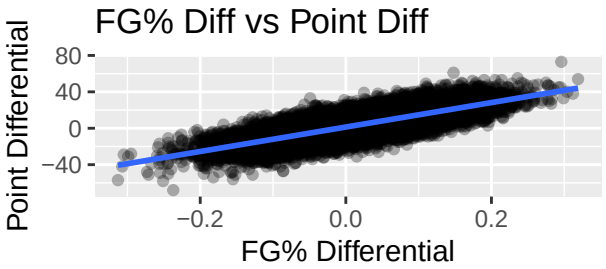


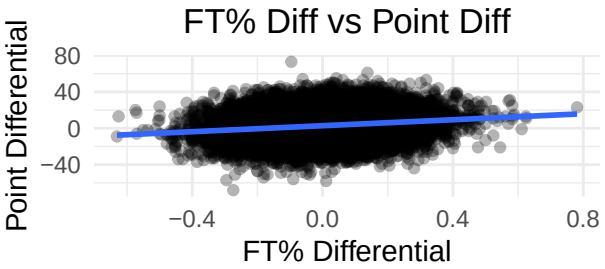
The boxplots above compare each differential performance metric between games where the home team won and games where the home team lost. Across all six statistics, home team wins are associated with higher differential values, meaning the home team generally outperformed the visiting team in these areas when they won.

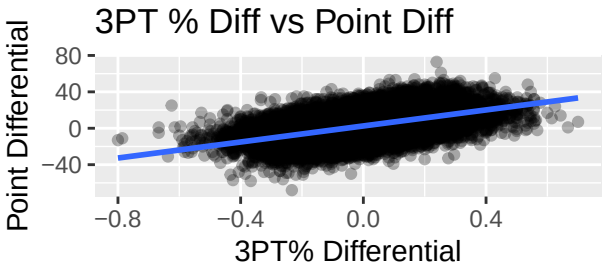
Assist and rebound differentials show large separation between win and loss outcomes, indicating that advantages in ball movement and possession are strongly related to winning. Shooting efficiency metrics (FG%, FT%, and 3PT%) display smaller but still meaningful shifts toward positive values in home wins. These patterns support the idea that both scoring efficiency and overall control of the game contribute to successful outcomes.

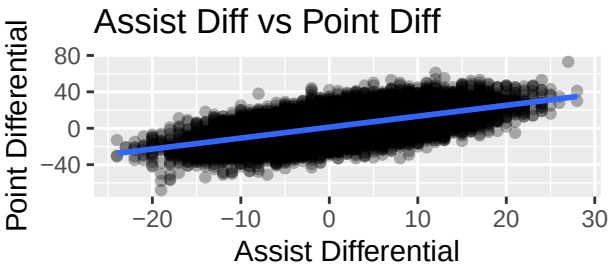
These visualizations give supple initial evidence that the selected predictors differ between wins and losses. This justifies their usage in the logistic regression model,

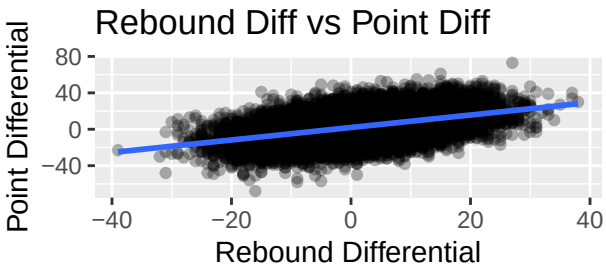
The scatter plots below represent relationships between point differential and the differential of five statistics. Comparing these is a good way to analyze the impact that each of the statistics can have on the outcome of a game. Every statistic, including point differential, is measured by the home teams number minus the away teams number. This way, everything is consistent. All of the plots have an upward facing slope, which fits with our intuitive expectations. Winning a differential of any of these statistics helps with winning and overall point differential. However, some have a much stronger impact on the outcome of the game than others.











- The FG% differential plot has the steepest slope among all metrics analyzed. When a team has a higher percentage of shots made, they will be more likely to win because they will be scoring more of their shot attempts. The larger the differential, the larger the point differential will be assuming that both teams take the same number of shots. It is important to note that winning the field goal % differential does not guarantee winning the game, as outside factors like turnovers and number of shots taken can affect winning in separate ways.
- The FT% differential plot is far less steep than the FG% differential plot, likely due to free throws not being a very large percentage of overall scoring. Additionally, the amount of total free throws attempts can be very different with certain teams due to style of play, officiating, and luck.
- The 3PT% differential plot follows a similar pattern to the field goal percentage graph, although it is not as steep. The 3PT% differential serves as a middle ground between FT% and FG%. Three pointers intuitively have a larger effect on the outcome of a game than free throws because of their intrinsic higher value (3 points $>$ 1 point) and the increasing number of 3 point shots taken in the modern NBA. 3PT% has a similar situation where the attempts can fluctuate, meaning that it is possible to have a negative point differential despite having a higher 3PT%.
- The assist differential plot exhibits a distinct shape compared to percentage-based metrics because assists are recorded as discrete counts rather than proportions. Since an assist is credited only when a pass directly leads to a made basket, higher assist totals generally correspond to more efficient offensive execution and, consequently, larger point differentials. Although not every scoring possession results from an assist, assist differential remains a meaningful performance indicator and shows a strong association with game outcomes.
- Similar to assists, rebounds are recorded as whole numbers, which contributes to the cleaner appearance of the rebound differential plot compared to percentage-based metrics. Rebounds are positively correlated with winning, though the relationship is weaker than that of assists. This is reasonable because rebounds do not directly generate points. Offensive rebounds can lead to second-chance scoring

opportunities, but the majority of rebounds in a game are defensive. While defensive rebounds do not add to a team's score, securing them prevents the opponent from extending possessions, reflects missed opponent shots, and increases the number of possessions for a team, all of which contribute indirectly to a stronger point differential.

Analysis

Table 1: Logistic Regression Coefficients for Predicting Home Team Win Probability

Term	Estimate	Std_Error	Z_value	P_value
(Intercept)	0.249	0.023	11.02	<2e-16
FG_diff	25.216	0.501	50.32	<2e-16
FT_diff	4.047	0.168	24.10	<2e-16
FG3_diff	4.643	0.172	26.99	<2e-16
AST_diff	0.074	0.004	17.69	<2e-16
REB_diff	0.091	0.003	31.62	<2e-16

The logistic regression model (win_model) evaluates how differential performance metrics influence the probability of a home team winning. Below is a detailed interpretation of each predictor:

- **Field Goal Percentage Differential (FG_diff)**: A coefficient of 25.352 indicates that for every 1 unit increase in field-goal percentage differential (home FG% - away FG%), the log-odds of the home team winning increase by approximately 25.35. This extremely large coefficient reflects the strong relationship between shooting efficiency and wins. The p-value is < 2e-16, demonstrating overwhelming statistical significance. This confirms that FG% differential is one of the strongest predictors of game outcome.
- **Free Throw Percentage Differential (FT_diff)**: With a coefficient of 4.176, a 1-unit increase in free throw percentage differential is associated with a 4.18 increase in the log-odds of the home team winning. While a weaker predictor compared to FG%, free throw differential contributes to predicting game outcomes, as shown by the p-value (< 2e-16).
- **Three Point Percentage Differential (FG3_diff**): With a coefficient of 4.591, a 1-unit increase in free throw percentage differential is associated with a 4.591 increase in the log-odds of the home team winning. With the p-value (<2e-16), 3PT% is a statistically significant predictor and reflects the importance of the 3 point shot for winning games in the modern NBA.
- **Assist Differential (AST_diff)**: The coefficient of 0.0766 indicates that each additional assist differential increases the log-odds of winning by 0.0766. While a statistically significant predictor because of the p value (<2e-16), it is the weakest predictor of winning games compared to the rest of the differential performance metrics, reflecting the shift away from team play in the modern NBA.
- **Rebound Differential (REB_diff)**: The coefficient of 0.0914 means that for every extra rebound differential, the log-odds of a home win increase by approximately 0.091. This variable is also statistically significant (p < 2e-16), but similar to the assist differential, its effect on predicting winning games is modest compared to shooting percentages. This implies that while rebounding contributes to game success, it is far less influential than shooting efficiency.

After fitting the logistic regression model, we generated predicted win probabilities for each game. For each observation, the model outputs the probability that the home team wins based on the statistical differentials. We then classify each game as a predicted home win or predicted away win using a 0.5 probability cutoff. The data used on our model was `test_data`. Our model had an 84.73% of predicting game outcomes correctly using `test_data`.

Table 2: Sample of Model Predictions for Game Outcomes

	Home_Team	Away_Team	Actual_Winner	Predicted_Winner	Predicted_Prob	Correct
3	Cavaliers	Bucks	Cavaliers	Bucks	0.436	FALSE
15	Heat	Bulls	Bulls	Bulls	0.055	TRUE
19	Cavaliers	Jazz	Cavaliers	Cavaliers	1.000	TRUE
28	Celtics	Magic	Magic	Magic	0.338	TRUE
31	Raptors	Warriors	Warriors	Warriors	0.024	TRUE
35	Clippers	Wizards	Clippers	Wizards	0.461	FALSE

Conclusion

This analysis demonstrates that differential team performance metrics are strong predictors of NBA game outcomes. Shooting statistics, especially field goal percentage differential and three-point percentage differential, were the most influential factors. These findings were supported consistently across our data visualizations that showed clear separation between winning and losing teams in shooting-related differentials, and by the logistic regression model, where shooting metrics produced the largest and most statistically significant coefficients.

Assists and rebounds also contributed to win probability, though their effects were comparatively much smaller. This suggests that while ball movement and rebounding support overall team success, they are less decisive than shooting efficiency in predicting outcomes in the modern NBA. Assists were the weakest contributing factor in our model in predicting outcomes which we believe is a direct reflection of the NBA moving away from team play.

The direction and magnitude of each coefficient aligned with expectations: teams that shoot better, pass better, and secure extra possessions via rebounds tend to outperform their opponents.

Overall, our logistic regression model successfully captured the relationship between winning NBA games and performance differentials. Although logistic regression provided strong baseline predictive insight, future improvements can include exploring non-linear models, incorporating player-level metrics, or account for variables such as rest days, travel distance, or injuries. Expanding the feature set would likely improve predictive accuracy and provide a more comprehensive understanding of the factors that drive NBA game results.