Core Metrics That Influence NBA Outcomes During the Regular Season

DSC 680 – Project 1, Milestone 3

Xhoi Shyti

Summary

As professional basketball becomes increasingly influenced by data-driven decision-making, teams are turning to statistical analysis to gain a competitive edge. This white paper presents the results of a comprehensive analysis of more than 70 years of NBA box score data, aimed at identifying which performance metrics are the most reliable indicators of a team's likelihood of winning or losing a game. Through a blend of exploratory data analysis, correlation analysis, and logistic regression, the research highlights shooting accuracy—particularly free throw percentage, field goal percentage, and three-point percentage—as the most influential factors in determining game results. These efficiency-based metrics proved more predictive of victories than total points scored or the number of shot attempts, underscoring the importance of precision rather than volume in offensive performance.

Problem Statement

In a league where games are frequently determined by the smallest of margins, identifying the statistics that most accurately predict victories is essential. Teams use this insight to enhance overall performance, shape effective game-day strategies, and support talent development and recruitment efforts. Additionally, these findings help drive informed, data-based decisions across all levels of the organization.

Methodology

Data Source

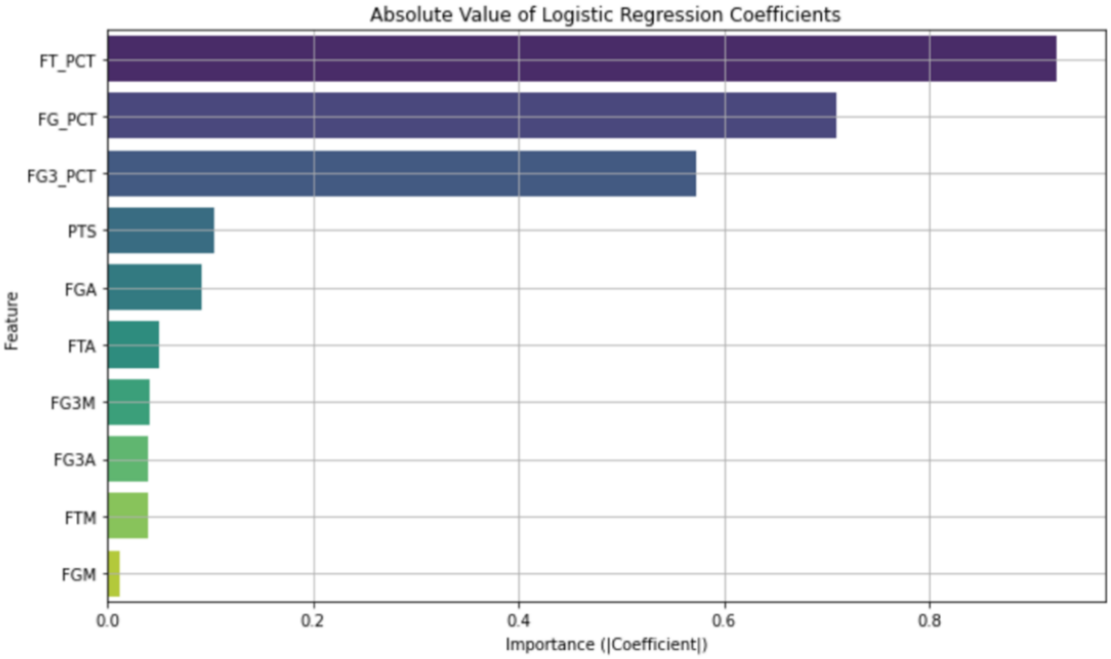
A custom Python script was used to collect historical NBA box score data spanning from 1946 to 2023 through the NBA Stats API. The dataset includes a wide range of game information, such as win-loss results, shooting performance, and key gameplay metrics like rebounds, assists, and turnovers.

Analytical Approach

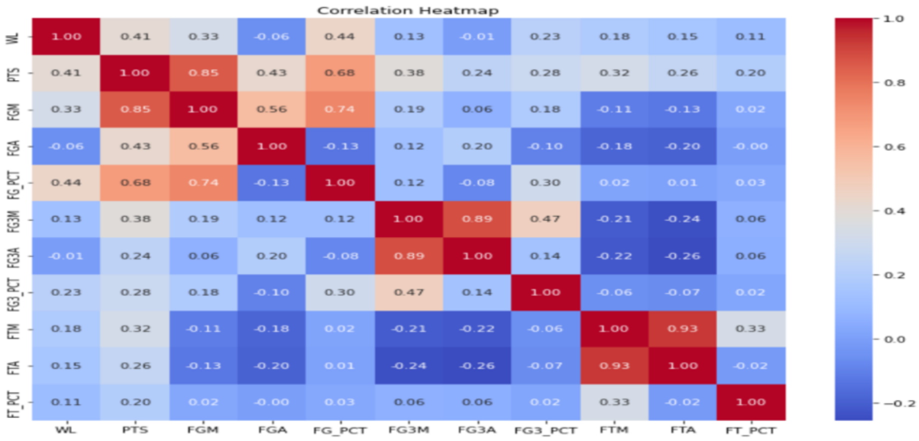
1. Exploratory Data Analysis: Revealed underlying trends and correlations within the dataset.
2. Feature Selection: Determined which metrics had the strongest statistical relevance.
3. Logistic Regression Modeling: Used key variables to estimate the likelihood of game outcomes.
4. Model Evaluation: Measured effectiveness using metrics such as accuracy, precision, and recall.

Key Findings

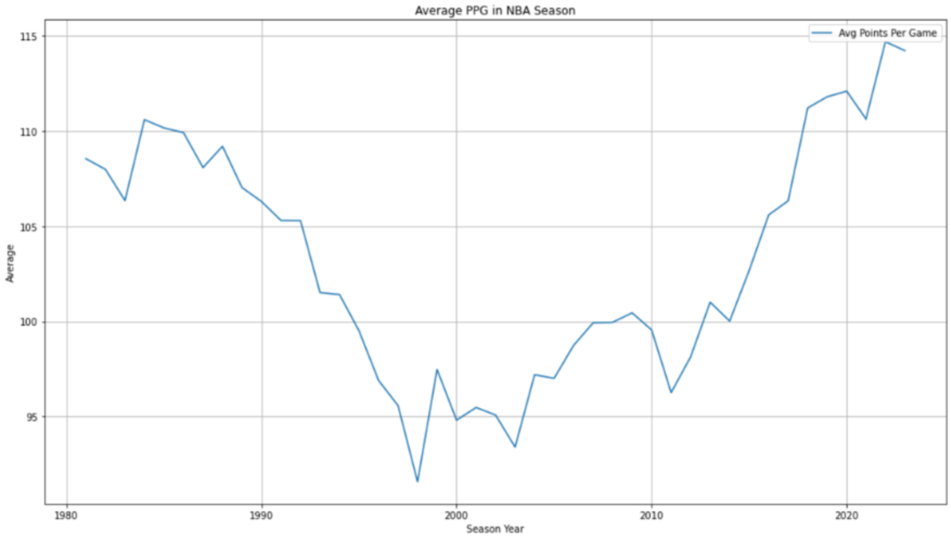
1. Shooting Accuracy Leads the Way
    - Free throw percentage emerged as the most influential factor, with field goal and three-point percentages closely behind.
    - Metrics based on scoring volume, such as total points, were less impactful than those reflecting shooting efficiency.
    - Teams should prioritize strategic shot selection and consistent free throw performance over simply increasing shot volume.



2. Key Findings from Exploratory Analysis
    - Heatmaps revealed strong correlations between made field goals, free throws, three-pointers, and total points scored.
    - Both field goal and free throw percentages showed a positive association with winning, highlighting the critical role of shooting efficiency in game outcomes.



-



-

3. Model Results
    - The model achieved an accuracy of approximately 71.5%.
    - Precision and recall were well-balanced for both winning and losing classifications.
    - Based on the confusion matrix, it correctly predicted 6,321 wins and 6,500 losses, with around 2,550 incorrect predictions in each category.

Assumptions & Limitations

The model has several limitations that should be considered when interpreting the results. First, it simplifies outcomes into a binary win or loss, without accounting for contextual factors such as point differentials, high-pressure moments, or the strength of the opposing team. Additionally, multicollinearity is present, as certain variables—like field goals made and field goal percentage—are mathematically related, which can complicate the interpretation of model coefficients. The dataset spans multiple decades, meaning shifts in rules and playing styles over time could lead to inconsistencies across different eras. Lastly, the analysis does not incorporate advanced statistics such as Player Efficiency Rating (PER), on/off court impact, or other modern performance metrics.

Ethical Considerations

It is important to maintain transparency and accuracy when collecting and interpreting data to ensure credible analysis. Conclusions should be based strictly on the evidence and should not make unfounded claims about specific players or teams. Additionally, it is essential to recognize that shifts in data collection practices over time may introduce historical biases that could affect the analysis.

Challenges Encountered

The analysis faces several challenges that must be addressed to ensure valid results. One major issue is the inconsistency of data collected over different decades, which can complicate comparisons across time periods. Additionally, the presence of multicollinearity among certain variables necessitates careful feature selection to avoid misleading interpretations. Interpreting raw statistical measures also proves difficult, as their meaning can vary depending on the evolving and situational nature of the game.

Strategic Implications

Organizations and coaching staff can apply these findings in several impactful ways. By emphasizing shooting efficiency in player development programs, teams can foster more effective scoring habits. Increased attention to free throw and three-point shooting during training can further boost performance. Additionally, integrating predictive modeling into strategic planning allows for more informed decision-making. These insights also enable a more targeted evaluation of both individual players and overall team performance.

Future Recommendations

Future research can be strengthened by integrating advanced statistics and player tracking data to provide deeper insights. Adjusting predictive models to account for game context—such as home versus away games or back-to-back scheduling—can improve accuracy. Exploring non-linear modeling techniques like random forests or XGBoost may also uncover more complex patterns in the data. Additionally, conducting analyses tailored to specific teams or player positions could yield more targeted and actionable results.

Audience Questions and Answers

1. What motivated the choice of logistic regression instead of using more advanced models such as random forests or neural networks?
    a. Logistic regression was chosen due to its clarity, ease of use, and strong performance in binary outcome prediction. In the context of this project—where both educational value and strategic insights were key—being able to clearly understand and explain the influence of each variable was essential. Unlike more complex models such as random forests or neural networks, logistic regression offers straightforward coefficients that help identify which factors most significantly impact game results. This level of transparency is especially valuable when presenting insights to stakeholders without technical backgrounds, such as coaches or team executives.
2. How was the issue of multicollinearity among closely related features like field goals made, field goals attempted, and shooting percentage handled?
    a. To manage multicollinearity, the analysis favored efficiency-focused variables instead of raw counts. For instance, field goal percentage was kept in the model, while field goals made and attempted were excluded, as FG% reflects shooting performance without duplicating information. Additionally,

a correlation matrix helped identify variables with high overlap, allowing for their removal or consolidation prior to training the model.

3. What methods were used to ensure the model remains applicable to future games or across different periods in NBA history?
    a. With data covering more than 75 years of NBA history, the model benefits from strong potential for generalization. To reduce the risk of overfitting to trends specific to certain time periods, the analysis focused on core statistics—such as shooting accuracy—that maintain their relevance across different eras of play. Furthermore, regularization techniques were considered to help minimize the influence of era-dependent variations and ensure a more stable model.
4. Why did free throw percentage emerge as the most influential metric, and is this trend consistent across various teams and historical eras?
    a. Free throw percentage represents a consistent and isolated scoring opportunity, unaffected by defensive pressure, making it a reliable performance indicator. Its relevance has remained stable across different team strategies and historical eras. Close contests are frequently influenced by free throw performance, particularly during late-game situations. Although there may be some variation between teams, FT% has consistently demonstrated a strong association with winning outcomes throughout both early and contemporary NBA seasons.
5. In what ways might these results influence how teams approach player development and roster construction?
    a. Organizations are encouraged to focus on identifying and nurturing players who demonstrate high shooting accuracy, especially in free throw situations. This approach could lead to a shift in scouting priorities, favoring players known for their efficiency rather than sheer scoring volume. Additionally, training initiatives may benefit from placing increased emphasis on enhancing free throw and three-point shooting skills across all roles on the court.
6. Did you observe any differences in the model’s predictive strength when separating home versus away games or comparing regular season to playoff matchups?
    a. The current model does not account for contextual factors within games. Nonetheless, it's likely that elements such as home court advantage or playoff intensity can impact player performance and decision-making. Future versions of the model could incorporate categorical variables representing game setting and type to better capture how these situational factors affect predictive accuracy.
7. How might the exclusion of advanced metrics such as Player Efficiency Rating or usage rate impact the model’s overall accuracy?
    a. Leaving out advanced statistics may reduce the model’s overall accuracy, as such metrics offer more detailed insights into player efficiency and involvement. However, the primary aim of this model was to maintain clarity and ensure wide applicability, which is better supported by standard box

score data. Future enhancements that incorporate metrics like Player Efficiency Rating, usage rate, or player tracking information could lead to improved predictive performance.

8. Given the model is based on binary win/loss outcomes, how would it need to be adjusted to predict point differentials or game margins?
    a. Predicting point differentials would be more effectively handled by a different type of model, such as linear regression. Another option could involve using ordinal or multiclass classification to categorize outcomes into ranges, like narrow losses, close wins, or blowouts. While this binary classification model offers a useful starting point, estimating point spread demands a distinct modeling strategy tailored to continuous or grouped outcomes.
9. What adaptations would be required to apply this model to individual player performance or fantasy basketball projections?
    a. Modifying the model for player-focused analysis would require incorporating individual-level statistics, such as shooting efficiency and on/off court differentials, along with game-by-game performance data. The prediction target could shift from team victories to outcomes like fantasy points or contributions to net rating. While logistic regression could remain a viable method, the model would need to integrate player-specific metrics and their interactions to capture individual impact accurately.
10. What future enhancements do you plan to make to the model, both from an analytical standpoint and in practical application?
    a. From an analytical perspective, future improvements would involve integrating advanced metrics such as Player Efficiency Rating (PER), True Shooting Percentage (TS%), and SportVU movement data to capture a more nuanced view of player performance. I would also test ensemble methods like Random Forest and XGBoost to enhance predictive accuracy, while incorporating contextual variables such as game location, opponent strength, and pace of play. On the practical side, the model could be implemented within an interactive dashboard to support coaching and scouting decisions. Outputs would be designed to offer clear, actionable insights—for instance, estimating how a 5% increase in free throw percentage affects a team's chances of winning. Additionally, the model could be explored for real-time applications, helping inform strategic decisions during games.