Key Metrics for Regular-Season NBA Games

DSC 680 – Project 1, Milestone 2

Xhoi Shyti

Summary

In the era of data-driven sports strategy, professional basketball teams increasingly seek to transform raw statistics into winning strategies. This white paper summarizes findings from an in-depth study that leveraged over seven decades of NBA box score data to answer a central question: What performance metrics most strongly predict whether an NBA team will win or lose a game?

Using a combination of exploratory data analysis (EDA), correlation mapping, and logistic regression modeling, the study identifies shooting efficiency—especially free throw percentage (FT%), field goal percentage (FG%), and three-point percentage (3P%)—as the strongest predictors of game outcomes. These metrics outperformed raw scoring totals and shot volumes in predicting wins, emphasizing quality over quantity in offensive execution.

Problem Statement

With games often decided by razor-thin margins, understanding which statistics best forecast wins is crucial. Teams aim to use this knowledge to:

- Optimize performance
- Inform game-day strategies
- Guide talent development and acquisition
- Make data-driven decisions at all organizational levels

Methodology

Data Source

Historical NBA box scores (1946–2023) were compiled via a custom Python script using the NBA Stats API. The data includes:

- Game outcomes (Win/Loss)
- Shooting statistics
- Rebounds, assists, turnovers, etc.

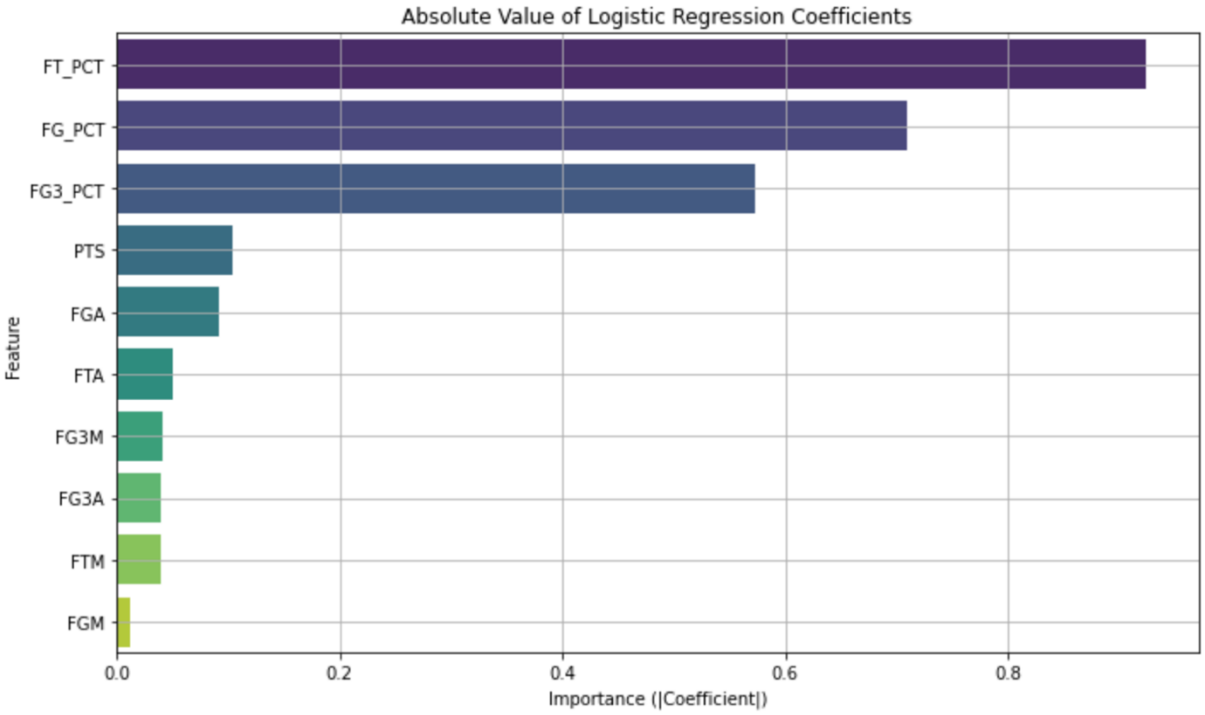
Analytical Approach

1. Exploratory Data Analysis: Uncovered patterns and statistical relationships.
2. Feature Selection: Identified statistically significant metrics.
3. Logistic Regression Modeling: Predicted game outcomes using selected features.
4. Performance Evaluation: Assessed model using accuracy, precision, recall.

Key Findings

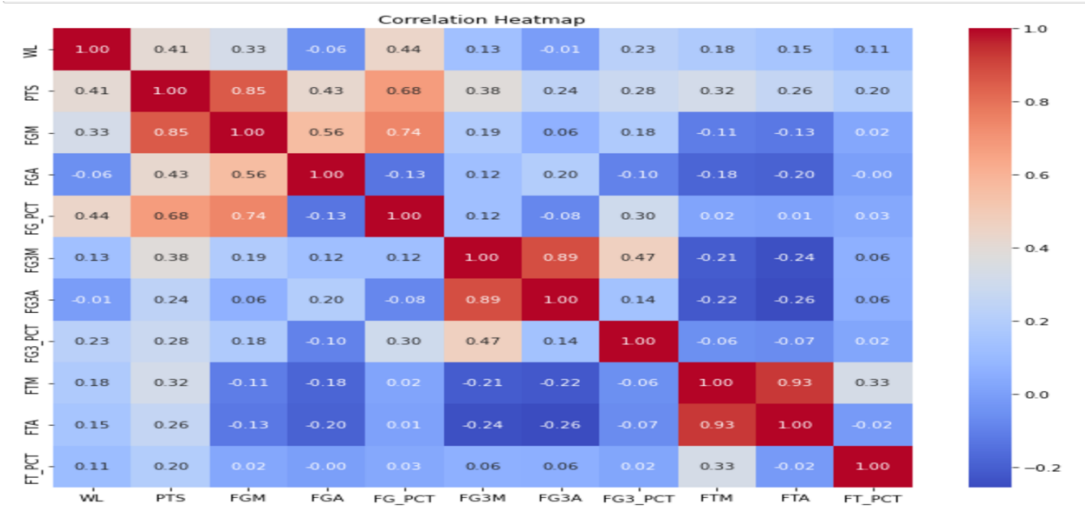
1. Shooting Efficiency Dominates

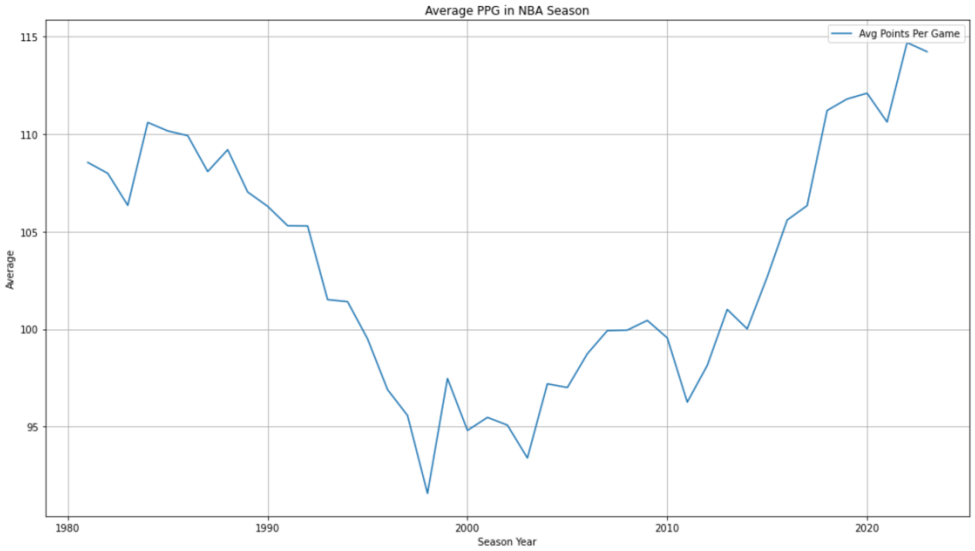
- FT% was the top predictor, followed by FG% and 3P%.
- Pure volume stats like total points (PTS) contributed less than efficiency-based stats.
- Emphasis should shift toward quality shot selection and free throw reliability.



2. EDA Insights

- Correlation heatmaps confirmed strong links between FGM, FTM, FG3M, and PTS.
- FG% and FT% positively correlated with wins, reinforcing the importance of conversion rates.





3. Model Performance

- Accuracy: ~71.5%
- Balanced precision/recall across both Win and Loss outcomes
- Confusion Matrix:
    - Correct Wins: 6,321
    - Correct Losses: 6,500
    - Misclassifications: ~2,550 each

Assumptions & Limitations

- Binary Simplification: Model does not consider context like margin of victory, clutch moments, or opponent strength.
- Multicollinearity: Some features (e.g., FGM and FG%) are mathematically related, affecting coefficient interpretation.
- Era-Spanning Data: Changes in rules and styles may introduce historical inconsistencies.
- No Advanced Metrics: Model excludes PER, on/off splits, and other advanced analytics.

Ethical Considerations

- Ensure transparency and accuracy in data sourcing and interpretation.
- Avoid drawing unjustified conclusions about individual players or teams.
- Acknowledge historical bias due to changes in data collection standards over time.

Challenges Encountered

- Data inconsistencies across decades
- Multicollinearity requiring feature selection techniques
- Interpreting raw stats in dynamic, situational contexts

Strategic Implications

Teams and front offices can leverage these insights to:

- Prioritize efficient shooting in player development
- Allocate more training to free throw and 3-point shooting
- Enhance game plans with predictive modeling
- Evaluate player and team performance through a more focused lens

Future Recommendations

- Incorporate advanced stats and player tracking data
- Adjust models for game context (home vs away, back-to-back games)
- Explore non-linear models like random forests or XGBoost
- Conduct team-specific or position-specific predictive analyses

Audience Questions

1. Why did you choose logistic regression over more complex machine learning models like random forests or neural networks?
2. How did you address multicollinearity between related variables like FGM, FGA, and FG%?
3. What steps did you take to ensure the model is generalizable to future games or different NBA eras?
4. Why is free throw percentage the most predictive factor—does this hold true across all teams or eras?
5. How do these findings change the way a team should approach roster building or player development?
6. Are there differences in predictive power when isolating home vs. away games or playoff vs. regular season?

7. How does the exclusion of advanced stats (like PER or usage rate) affect the accuracy of your model?
8. Since your model uses binary outcomes (Win/Loss), how would it handle predicting point spreads or margins of victory?
9. How could you tailor this model for individual player contributions or fantasy basketball projections?
10. What would be your next steps to enhance this model—both analytically and practically?