## Abstract

This study leverages advanced statistical and machine learning techniques to analyze basketball performance data. Using a comprehensive dataset, we computed rolling averages for key performance metrics, aggregated player-level statistics to team-level features, and employed feature selection techniques such as LASSO regression and PCA for dimensionality reduction. The analysis reveals significant factors influencing game outcomes and provides a framework for predictive modeling. Limitations include data quality issues and potential biases in feature engineering, which suggest avenues for future research.

## Introduction

Understanding the factors that contribute to basketball game outcomes is critical for teams, analysts, and enthusiasts. This paper explores how rolling averages, aggregated player-level statistics, and advanced feature selection techniques can uncover key insights. Using a comprehensive dataset, which spans 1320 games, we employed methods such as LASSO regression and PCA to identify the most influential features while reducing dimensionality. Key variables in the analysis include points scored, effective field goal percentage, turnovers, and true shooting percentage. By integrating short-term performance trends and aggregated player-level data, this study offers a robust framework for predictive analytics in sports.

The findings from this study highlight critical performance metrics and provide actionable insights for teams aiming to optimize strategies and improve outcomes. The structure of the paper includes a discussion of the data and preprocessing steps, an outline of the analytical methodologies, a presentation of the results, and a conclusion that explores the implications, limitations, and potential areas for future research.

## Data

The dataset includes game-level and player-level statistics from multiple basketball seasons, capturing both basic and advanced metrics. Game-level data consists of 1320 games with variables such as points scored (mean of 115.63 for home teams and 113.03 for away teams), field goal attempts (mean of 88.18 for home teams), and turnovers. Advanced metrics include effective field goal percentage (mean of 0.540 for away teams), true shooting percentage (mean of 0.576 for away teams), and pace (mean of 99.60 for away teams).

Preprocessing steps involved cleaning the data by addressing missing values and inconsistent formats. For instance, timestamps in the dataset were converted into a standardized datetime format to ensure accurate chronological ordering. Outliers were identified and reviewed for potential impact on the analysis. Player-level data, consisting of detailed performance metrics such as usage percentage and rebound chances, was aggregated to create team-level features.

## Methods

Rolling averages were computed to capture short-term performance trends. Metrics such as points, rebounds, and shooting percentages were averaged over the last five games, enabling the analysis to reflect recent form. This required the data to be sorted chronologically by team and game date. Aggregating player-level data involved calculating mean values for key statistics, ensuring a holistic representation of team performance. Metrics such as points off turnovers, second-chance points, and effective field goal percentage were included in the aggregated features.

To identify significant predictors of game outcomes, LASSO regression was applied. This method shrinks less relevant variables to zero, leaving only the most influential features. In addition, PCA was used to reduce dimensionality while retaining 95% of the variance, enhancing model interpretability and computational efficiency. Features were standardized before applying these techniques to ensure consistency in scale.

## Results

The analysis uncovered several significant findings. LASSO regression identified points per game, effective field goal percentage, and turnovers as critical factors in determining game outcomes. Teams with consistent performance in these metrics over recent games, as reflected by rolling averages, exhibited stronger predictive signals. Dimensionality reduction through PCA reduced the feature set to a manageable subset, retaining essential information while improving computational efficiency. The cumulative variance explained by the first few principal components highlighted the compactness of the feature space.

Visualizations generated in the analysis included feature importance plots from the LASSO model and explained variance ratios from PCA. These plots provided clarity on the relative importance of different metrics and the effectiveness of dimensionality reduction.

## Conclusion

This study demonstrates the efficacy of combining traditional statistical methods with machine learning techniques to analyze basketball performance data. The findings emphasize the importance of metrics such as shooting efficiency and turnovers in predicting game outcomes.

While the analysis provides robust insights, limitations include potential biases in feature engineering and the sparsity of certain advanced metrics.

Future research could address these limitations by incorporating contextual factors such as player injuries and game locations. Additionally, enhancing the dataset with real-time performance metrics and exploring advanced predictive models such as neural networks could further refine the analysis. By turning the limitations of this study into opportunities for future work, this paper lays the groundwork for more sophisticated analyses in the field of sports analytics.

## References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python.

Basketball Reference. https://www.basketball-reference.com/