Beyond the Scoreboard: Advanced Machine Learning Models for Predicting NBA Game Outcomes and Team Performance

1st Aneri Patel 2nd Shrihan Thokala 3rd Samiksha Burkul 4th Swathi Kanneti Ramana Reddy

***Abstract*—Our Study employs machine learning to predict outcomes in NBA games using a dataset divided into five files. Our approach began with exploratory data analysis (EDA) to uncover patterns in game statistics and player performance. This led to the development of two models: the first predicts the home team's total score, integrating factors like player stats and team dynamics; the second, a binary classification model, forecasts the home team's likelihood of winning. These models blend accuracy with interpretability, serving as valuable tools for sports analysts and enthusiasts. The study not only offers insights into NBA game strategies but also contributes to sports data science, demonstrating the efficacy of machine learning in sports analytics. The outcomes suggest potential applications in team strategy development and betting markets, while also highlighting areas for future research in sports modeling.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

In the evolving landscape of professional sports, the integration of sports analytics has become crucial in shaping strategic decisions, particularly in the NBA. Advanced data gathering and analysis methods have significantly enhanced our understanding of player performance and game dynamics. This transition from traditional observation to a data-driven, quantitative approach has empowered teams and coaches to make more informed decisions, impacting game outcomes and player management.

Predicting NBA game outcomes presents a complex challenge due to variables like player performance, team dynamics, and home-court factors. Our study aims to address this by developing two machine learning models. The first model will predict the home team's total score by analyzing data points such as player statistics and team performance history. The second model will classify the likelihood of the home team winning, taking into account factors beyond basic statistics.

Our approach begins with an in-depth Exploratory Data Analysis (EDA) of our NBA dataset to identify key trends and relationships. Insights from this EDA will guide the development of our predictive models. Through this research, we aim to provide valuable insights into NBA game strategies and contribute to the broader field of sports analytics.

II. Methododlogy

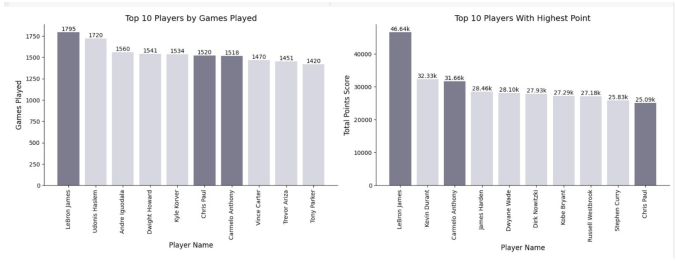
A. Data Set Overview

In this section of our analysis, we utilized several key datasets. The game statistics provided an overview of each match, detailing aspects like game IDs, dates, and participating teams. For more in-depth analysis, we employed detailed game statistics, which included player-specific metrics such as points scored, assists, and rebounds. Complementing this, player data was crucial, offering insights through player IDs, names, team affiliations, and personal attributes like height and weight. Finally, team data enriched our understanding of the broader context, encompassing team IDs, names, and additional relevant information. This comprehensive data collection allowed for an extensive examination of both team dynamics and individual player performances, vital for a thorough understanding of the games.

B. Data Pre-Processing

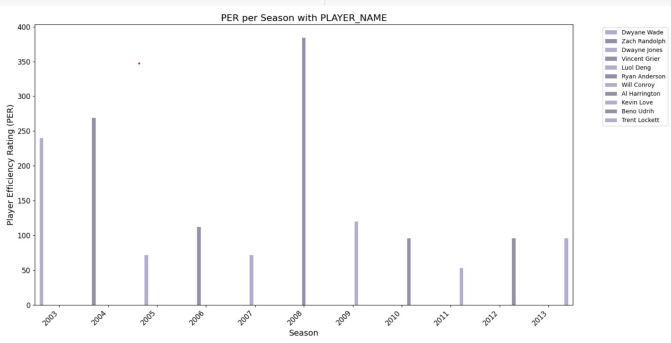
In this stage of our NBA games analysis, we undertook several crucial steps to refine our dataset for effective analysis. Initially, we focused on renaming the columns to ensure they accurately reflected the data they represented, thus enhancing the clarity and manageability of our dataset. We then proceeded to carefully select potential predictors along with our target variable, '`PTS_home`', to tailor our dataset specifically for the analysis. This involved creating a new, streamlined dataframe that included only the relevant columns. Furthermore, we conducted a correlation analysis by calculating a correlation matrix. This was a key step in our preprocessing, as it helped us identify variables with significant relationships to our target, thereby laying the groundwork for a more informed and effective modeling phase.

C. Exploratory Data Analysis

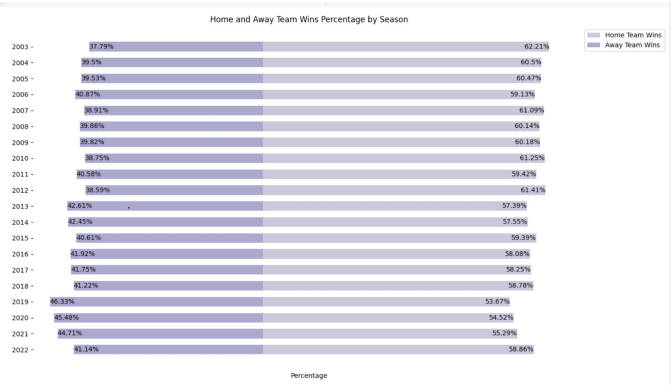


- LeBron James leads both categories, indicating not only a long and durable career but also a high level of performance throughout, as evidenced by the total points scored.

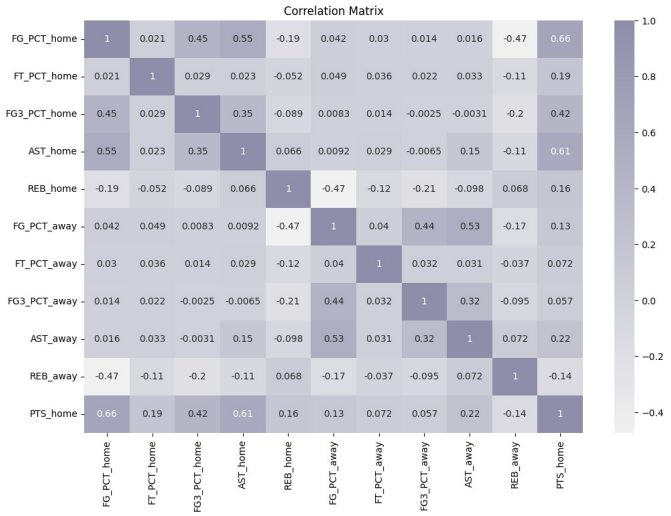
- There is not a direct correlation between the number of games played and the total points scored among these players, highlighting that scoring efficiency and opportunities differ from player to player.
- Vince Carter, while not among the top scorers, has one of the highest games played, suggesting a career with considerable longevity but perhaps with less scoring per game compared to some peers like Kevin Durant.



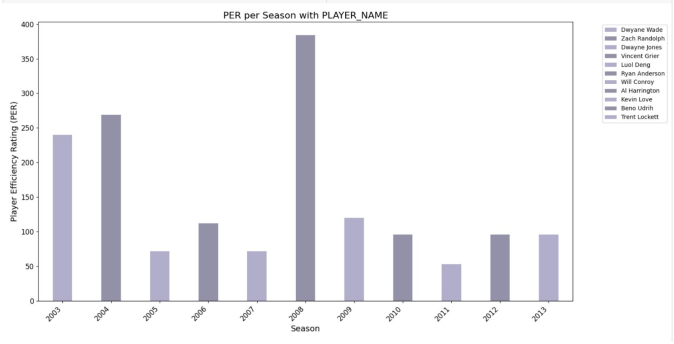
- The PER experiences a significant spike in 2007, which is an outlier compared to the other seasons, suggesting an extraordinary performance or a possible data error.
- There is considerable variability in PER from year to year, which may reflect changes in player performance, team dynamics, or other seasonal factors.
- The general pattern does not show a consistent improvement or decline over the seasons, indicating that the player's efficiency rating is subject to fluctuations possibly due to injuries, transfers, or changes in team strategy.



- Home teams consistently have a higher win percentage compared to away teams over the 20-year span, emphasizing the home-court advantage in the NBA.
- The gap between home and away win percentages appears to be narrowing in more recent seasons, particularly from 2019 to 2022, suggesting a possible increase in the competitiveness of away teams or a decrease in the traditional home advantage.
- The 2019 season shows the highest away team win percentage, which is significantly closer to the home team win percentage compared to other seasons, indicating a unique trend or influencing factors during that season that may have leveled the playing field.

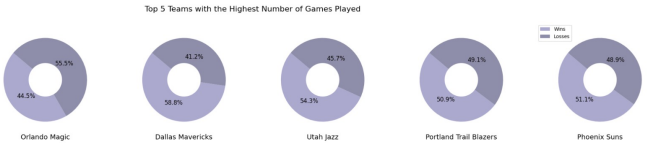


- Assists at home (AST_home) have a strong positive correlation with points scored at home (PTS_home), suggesting that team play characterized by higher assists tends to increase scoring efficiency.
- Field goal percentage at home (FG_PCT_home) is moderately correlated with three-point percentage at home (FG3_PCT_home) and points scored at home (PTS_home), indicating that overall shooting efficiency contributes to both perimeter shooting and total points.
- Rebounds at home (REB_home) have a weak to no correlation with field goal percentage away (FG_PCT_away), free throw percentage away (FT_PCT_away), and three-point percentage away (FG3_PCT_away), implying that rebounding performance at home does not significantly relate to shooting performance in away games.



- There is a notable peak in the PER in the 2007 season, suggesting an exceptional performance year for the player listed or a possible data outlier that warrants further investigation.
- The player's performance, as indicated by the PER, shows significant fluctuation over the years, with no clear upward or downward trend.

- The years following the peak in 2007 show a general decline in PER, with moderate recoveries in 2009, 2011, and 2013, indicating possible periods of comeback or improved performance.



The figure illustrates donut charts of the win-loss records for the top 5 NBA teams with the most games played. The Orlando Magic have a win percentage of 44.5% against a loss percentage of 55.5%, the Dallas Mavericks exhibit a more favorable outcome with 58.8% wins to 41.2% losses, the Utah Jazz have secured 54.3% of their games versus 45.7% losses, the Portland Trail Blazers display a nearly balanced record with 50.9% wins and 49.1% losses, and the Phoenix Suns hold a marginal majority of wins at 51.1% in contrast to 48.9% losses. Each chart employs a darker hue to signify the win percentage and a lighter hue for the loss percentage, showcasing a competitive balance among the teams, with the Mavericks notably achieving a higher win ratio.

D. Data Modeling

In the Data Modelling section of our methodology, we employed a diverse array of machine learning models to analyze the NBA dataset, concentrating on two main objectives: predicting the scores of the home teams based on various game statistics and classifying whether the home team would emerge victorious.

- **Home Team Score Predictor:**
  - **Linear Regression:** As the foundational model of our analysis, linear regression established a baseline for predictions. This model was trained on key game statistics such as shooting percentages (field goal, free throw, three-point) and rebound numbers, using standard regression techniques to elucidate the relationship between these independent variables and the dependent variable, home team scores. The model's performance was evaluated with metrics like R-squared, Mean Squared Error (MSE), and Mean Absolute Error (MAE) to ensure accuracy and reliability.

  - **Polynomial Linear Regression:** To address the limitations of linear models in capturing complex, non-linear relationships in the data, we applied a polynomial transformation to our linear regression model. This enhanced the model's ability to model the intricacies inherent in sports data and was particularly effective in identifying and leveraging subtle, non-linear patterns, crucial for a nuanced understanding of game outcomes.

  - **Random Forest Regression:** Chosen for its robustness in handling non-linearities and interactions between variables, Random Forest, an ensemble method that combines multiple decision trees, was used to produce more accurate and stable predictions. Excellently suited for our comprehensive NBA dataset, which contains a large number of features, the model's efficacy was assessed using R-squared and other relevant metrics.

  - **Support Vector Machine (SVM):** An SVM model for regression (SVR) was incorporated to model various aspects of the data, known for its effectiveness in high-dimensional spaces. The versatility of the model's kernel functions was key in capturing complex relationships in the dataset, and its performance was rigorously tested for accuracy and reliability.

  - **XGBoost:** We utilized XGBoost for its advanced capabilities in managing large and complex datasets. Known for its efficiency and scalability, this gradient boosting framework was tuned for our specific dataset, focusing on optimizing parameters like learning rate, depth of trees, and regularization. The model's performance was thoroughly validated using metrics such as R-squared, MSE, and RMSE (Root Mean Squared Error), ensuring its robustness in predicting game outcomes.

- **Home Team Win Classifier:**
  - **Logistic Regression:** Our study began with Logistic Regression, focusing on predicting the likelihood of the home team's victory using game statistics. The data was split into training and testing sets and standardized with StandardScaler. We trained the Logistic Regression model on this scaled data. Its performance was evaluated using metrics like accuracy score, ROC curve, and AUC score, with the ROC curve plotting the true positive rate against the false positive rate, and the AUC score quantifying the model's discriminatory power.

  - **Support Vector Machine (SVM):** The SVM model was utilized to classify game outcomes (win or lose for the home team) using game statistics. We trained the model on standardized features, ensuring effectiveness in handling the dataset's high-dimensional space. The model's accuracy, along with a confusion matrix and

classification report, provided a comprehensive evaluation of its predictive performance.

* **K-Nearest Neighbors (KNN):** We employed the K-Nearest Neighbors model to predict home team victories based on selected game statistics. The KNN model was developed using a scaled feature set, with the number of neighbors optimized for performance. The model's accuracy was assessed, and we used a confusion matrix and classification report for an in-depth analysis of its predictive capabilities.

* **Decision Tree:** Additionally, a Decision Tree model was included in our methodology. Decision Trees are particularly useful for their simplicity and interpretability, offering a straightforward visual representation of the decision-making process. We trained the Decision Tree on our dataset to identify patterns and factors influencing the home team's chances of winning. The model's performance was measured using accuracy, and we examined the tree's structure to understand the key determinants in game outcomes.

RESULTS I. Home Team Score Predictors a) Linear Regression :
These metrics assess the performance of a predictive model. The MSE, RMSE, and MAE measure the differences between predicted values and actual values, with lower values indicating better predictive accuracy. The R-squared value, ranging from 0 to 1, indicates the proportion of variance in the dependent variable that's predictable from the independent variable(s). In this case, an R-squared value of 0.75 suggests that 75% of the variance in the dependent variable is explained by the independent variable(s) in the model.

```
Mean Squared Error: 43.569985782009816
Root Mean Squared Error: 6.600756455286759
Mean Absolute Error: 5.1835799898542065
R-squared: 0.7525821888986025
```

b) Polynomial Linear Regression: These metrics are common indicators used to evaluate the performance of predictive models. The MSE, RMSE, and MAE measure the average differences between predicted values and actual values, with lower values indicating better accuracy. The R-squared value, ranging from 0 to 1, represents the proportion of variance in the dependent variable that is predictable from the independent variable(s). In this case, an R-squared value of 0.77 suggests that 77% of the variance in the dependent variable is explained by the independent variable(s) in the model.

```
Mean Squared Error: 40.435511203163244
Root Mean Squared Error: 6.358892293722488
Mean Absolute Error: 5.015816230465073
R-squared: 0.772501587653443
```

c) Support Vector Machine Regressor : This R-squared value, which indicates the proportion of variance in the dependent variable explained by the independent variable(s) in the SVM regression model, is approximately 0.7511. It suggests that around 75.11% of the variance in the dependent variable is accounted for by the independent variable(s) utilized within the SVM regression model.

```
0.7510757701985349
```

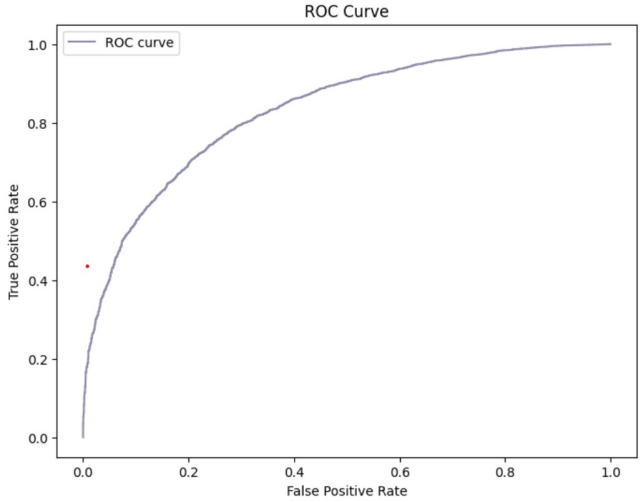
d) XG Boost : The R-squared (R^2) value of 0.6239 indicates that approximately 62.39% of the variability in the dependent variable is explained by the independent variables incorporated within the XGBoost regression model.

```
0.6238871402433316
```

II. Home Team Classifiers a) Logistic Regression: i) The Logistic Regression model achieved 76% accuracy in predicting home team outcomes in NBA games. It accurately predicted wins 78% of the time but had a 33% chance of misclassifying losses as wins. While it generally captured wins well (81% recall), it struggled with losses (67% recall), demonstrating balanced precision (72% for losses, 78% for wins) but slight inconsistency in recognizing losses. Overall, the model displays moderate predictive ability (weighted average F1-score 0.75) but might benefit from improved identification of losses to enhance overall performance.

```
Accuracy: 0.76

Confusion Matrix:
[[2202 1075]
 [ 870 3819]]

Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.67      0.69      3277
           1       0.78      0.81      0.80      4689

    accuracy                           0.76      7966
   macro avg       0.75      0.74      0.75      7966
weighted avg       0.75      0.76      0.75      7966
```

ii) ROC curve: With an Area Under the Curve (AUC) score of 0.83, the classifier exhibits commendable discriminative ability in distinguishing between positive and negative classes. However, the curve's inability to reach the top left corner implies potential for improvement in the model's performance, particularly in reducing the false positive rate while maintaining or increasing the true positive rate. This suggests avenues for refining the classifier's accuracy, emphasizing the need for fine-tuning to enhance its predictive capability.



AUC Score: 0.83

iii) C- parameter: The accuracy remained relatively stable across different C values, with the model consistently achieving a 76% accuracy in predicting the outcomes of basketball games. This suggests that the regularization parameter had a limited impact on the model's performance in this context. The choice of C did not substantially alter the accuracy, indicating that the logistic regression model, as configured with the default C value of 1, provided a robust and consistent predictive performance for the given dataset.

```
C=0.001, Accuracy: 0.75
C=0.01, Accuracy: 0.75
C=0.1, Accuracy: 0.76
C=1, Accuracy: 0.76
C=10, Accuracy: 0.76
C=100, Accuracy: 0.76
```

b) Decision Tree: The Decision Tree algorithm yielded a 67% accuracy in predicting home team outcomes. While the precision for predicting home team wins was higher at 72%, the model demonstrated a slightly lower overall accuracy compared to Logistic Regression. The decision tree's visualization of feature importance could offer valuable insights into the factors influencing game outcomes.

```
Decision Tree Accuracy: 0.67

Confusion Matrix:
[[1994 1283]
 [1326 3363]]

Classification Report:
              precision    recall  f1-score   support

           0       0.60      0.61      0.60      3277
           1       0.72      0.72      0.72      4689

    accuracy                           0.67      7966
   macro avg       0.66      0.66      0.66      7966
weighted avg       0.67      0.67      0.67      7966
```

c) Support Vector Machine (SVM): The Support Vector Machine achieved a commendable accuracy of 75%, striking a balance between precision and recall. Notably, SVM demonstrated strong performance in correctly predicting both home team victories (3903 instances) and losses (2111 instances). The precision for predicting home team wins was 77%, showcasing its efficacy in identifying positive outcomes.

```
SVM Accuracy: 0.75

Confusion Matrix:
[[2111 1166]
 [ 786 3903]]

Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.64      0.68      3277
           1       0.77      0.83      0.80      4689

    accuracy                           0.75      7966
   macro avg       0.75      0.74      0.74      7966
weighted avg       0.75      0.75      0.75      7966
```

d) k-Nearest Neighbors (KNN): The k-Nearest Neighbors algorithm attained an accuracy of 72%, positioning it as a competitive model in our analysis. With 3655 instances of correctly predicted home team wins, KNN showcased a precision of 75%. The balance between precision and recall suggests its effectiveness in capturing relevant patterns within the dataset.

E. CONCLUSION:

In conclusion, our study leveraged advanced machine learning models to predict NBA game outcomes and team

```
KNN Accuracy: 0.72

Confusion Matrix:
[[2081 1196]
 [1034 3655]]

Classification Report:
              precision    recall  f1-score   support

           0       0.67      0.64      0.65      3277
           1       0.75      0.78      0.77      4689

    accuracy                           0.72      7966
   macro avg       0.71      0.71      0.71      7966
weighted avg       0.72      0.72      0.72      7966
```

performance. The exploration began with a comprehensive Exploratory Data Analysis (EDA), revealing valuable insights into player performance, team dynamics, and game strategies. Two key models were developed: the Home Team Score Predictor and the Home Team Win Classifier.

For the Home Team Score Predictor, various models such as Linear Regression, Polynomial Linear Regression, Random Forest Regression, Support Vector Machine (SVM), and XGBoost were employed. Each model exhibited varying degrees of accuracy, with Polynomial Linear Regression leading with an R-squared value of 0.77, indicating that 77

The Home Team Win Classifier, focusing on predicting the likelihood of the home team winning, utilized Logistic Regression, Decision Tree, SVM, and K-Nearest Neighbors (KNN). Logistic Regression achieved 76

Insights gained from the EDA highlighted trends such as the importance of home-court advantage, the impact of player efficiency (PER), and correlations between assists, shooting percentages, and points scored. Additionally, the study presented donut charts illustrating win-loss records for the top 5 NBA teams with the most games played.

Our findings contribute not only to understanding NBA game dynamics but also showcase the potential applications of machine learning in sports analytics. The models developed can serve as valuable tools for sports analysts and enthusiasts, aiding in strategy development and decision-making. Furthermore, the study identifies areas for future research in sports modeling, emphasizing the continuous evolution of data-driven approaches in the realm of professional sports.

F. REFERENCES:

1. Cheng G, Zhang Z, Kyebambe MN, Kimbugwe N. Predicting the Outcome of NBA Playoffs Based on the Maximum Entropy Principle. Entropy. 2016; 18(12):450. https://doi.org/10.3390/e18120450

2.Chen WJ, Jhou MJ, Lee TS, Lu CJ. Hybrid Basketball Game Outcome Prediction Model by Integrating Data Mining Methods for the National Basketball Association. Entropy (Basel). 2021 Apr 17;23(4):477. doi: 10.3390/e23040477. PMID: 33920720; PMCID: PMC8073849.

3.Harmon, M. , Ebrahimi, A. , Lucey, P. , Klabjan, D. (2021). 'Predicting Shot Making in Basketball Learnt from Adversarial Multiagent Trajectories'. World Academy of Science, Engineering and Technology, Open Science Index 179, International Journal of Sport and Health Sciences, 15(11), 973 - 983.

4. E. Nalisnick. Predicting basketball shot outcomes. https://enalisnick. wordpress.com/2014/11/24/predicting-basketball-shot-outcomes/, 2014.

5. Loeffelholz, B., Bednar, E. Bauer, K. (2009). Predicting NBA Games Using Neural Networks. Journal of Quantitative Analysis in Sports, 5(1). https://doi.org/10.2202/1559-0410.1156

6. Papageorgiou, G. (2022). Data Mining in Sports: Daily NBA Player Performance Prediction.

7. Radovanović, S. (2016). Two-phased DEA-MLA approach for predicting efficiency of NBA players. Yugoslav Journal of Operations Research, 24(3).

8. Ma, B., Wang, Y., Li, Z. Application of data mining in basketball statistics. Applied Mathematics and Nonlinear Sciences.