University of Texas at Arlington



Applied Statistics and Data Science

5302 –Principles of Data Science

Final Project Report

**Submitted By**

**Group 7**

**The Classification Pack**

* Phuong Trinh
* Swathi Gutta
* Prithwiraj Chatterjee
* Pradhyumna Danduboina

**NBA Championship Prediction**

1. **Abstract:**

This project aims to predict the outcome of the NBA 2024–2025 championship using a combination of historical and real-time basketball data accessed via the NBA API. We collected, cleaned, engineered, and modeled team and player statistics using machine learning to identify which teams are most likely to win the championship. The study compares model performances using accuracy, F1-score, precision, and ROC AUC, with a final ensemble ranking and playoff prediction framework.

1. **Introduction:**

Basketball analytics has grown significantly over the years, transforming how fans, teams, and analysts interpret the game. The goal of this project is to harness the power of machine learning to predict which team is most likely to win the NBA championship for the 2024–2025 season. Leveraging historical performance, player statistics, and team metrics from the NBA API, this project will attempt to identify the key features influencing championship outcomes and build predictive models to make data-driven forecasts.

1. **Datasets:**

An API (Application Programming Interface) is a set of tools that allows different software systems to communicate and exchange data efficiently. For this project, we used the NBA API, an official source that provides real-time and historical basketball data in JSON format. These JSON files were parsed into structured Pandas DataFrames for analysis. The datasets included multiple endpoints:

* **leagueStandings**: 30 teams × 81 variables covering wins, losses, situational splits, and monthly performance
* **leagueDashTeamStats**: 30 teams × 54 raw features, filtered to ~14 essential ones (points, rebounds, etc.)
* **leagueDashPlayerStats**: 572 players × 66 metrics including per-game and advanced statistics

These endpoints were merged using TEAM\_ID and cleaned to form the final dataset. The dataset includes features such as Team ID, Points Scored, Assists, Rebounds, Turnovers, Win/Loss status, Field Goal Percentage, Home/Away games, Opponent performance, and Playoff standing. This combination of team-level and player-level metrics forms a robust foundation for predictive modeling.

1. **Data Preprocessing:**

Before modeling, we conducted thorough data cleaning and preprocessing to ensure high-quality input. The dataset contained no missing or duplicate values, which simplified the cleaning process. While a few mild outliers were observed in the boxplots of features like rebounds, plus-minus, or turnovers, they were not significant enough to warrant removal or transformation.

A group of blue squares

AI-generated content may be incorrect.

All numeric features were then normalized using standard scaling techniques to ensure that variables with different ranges did not disproportionately influence model training. This normalization step was especially important for distance-based algorithms and clustering. The resulting dataset was clean, standardized, and ready for modeling.

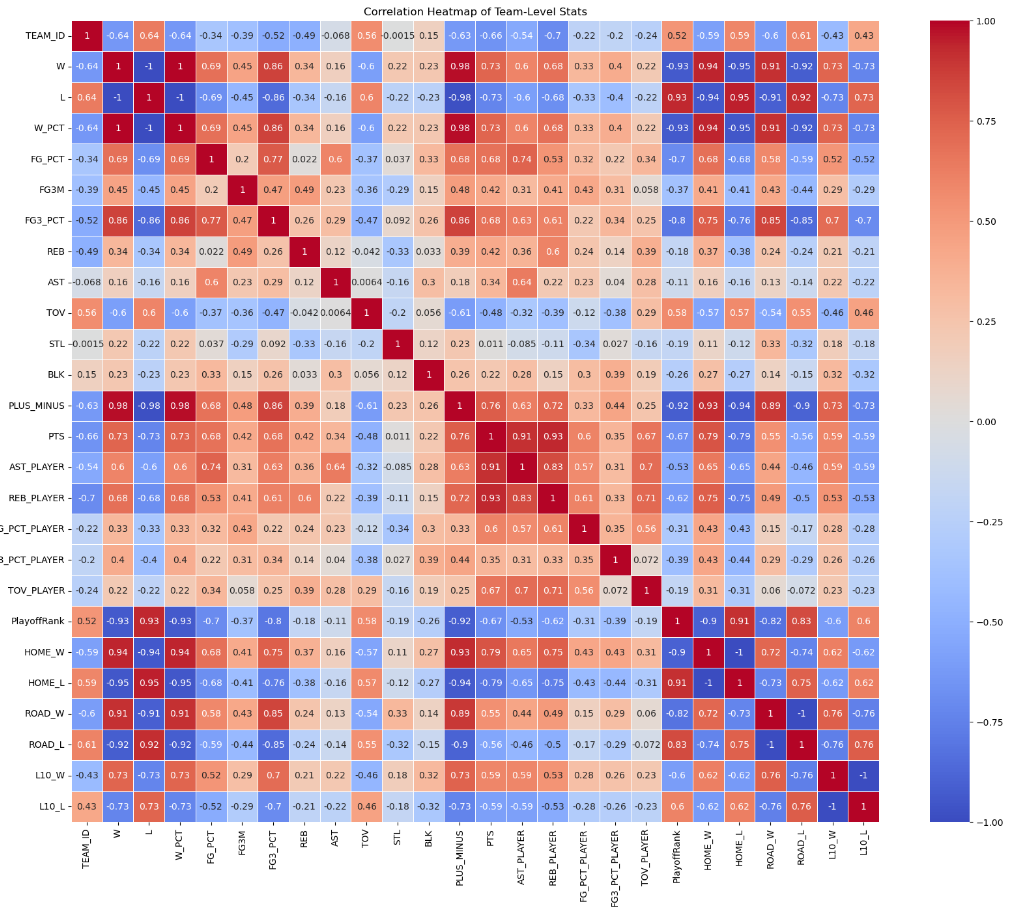
1. **EDA**

To better understand the structure and relationships within our NBA team dataset, we conducted an Exploratory Data Analysis focusing on statistical distributions, performance patterns, and inter-feature correlations. The dataset included a blend of raw statistics (e.g., Points, Rebounds, Turnovers), derived metrics (e.g., Win Percentage, Plus/Minus), and engineered player-level aggregates.

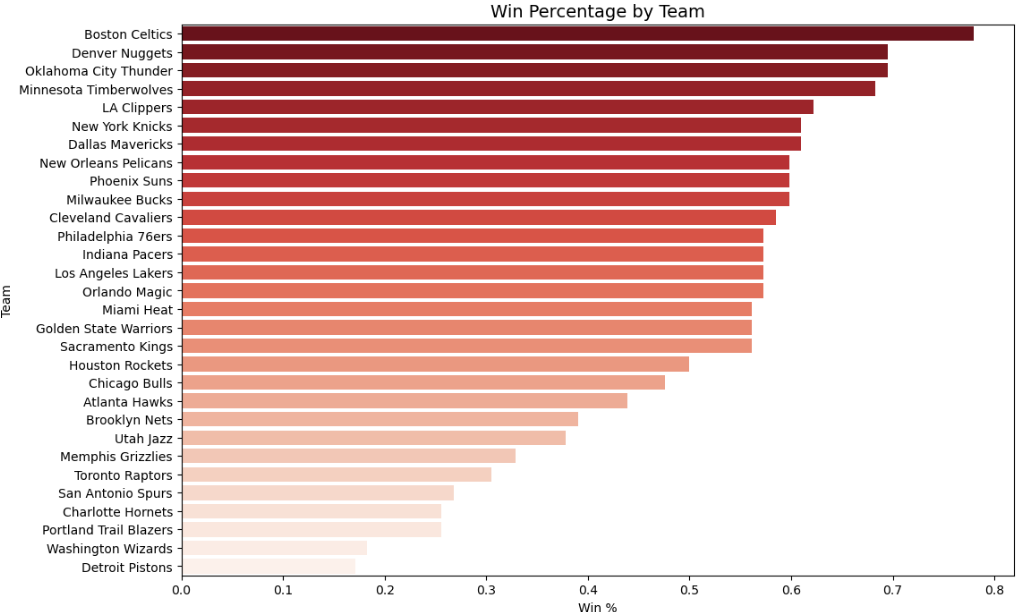
**1. Overview of Key Variables**

* W\_PCT (Win Percentage): Measures team success rate.
* PTS (Points): Total team scoring output.
* AST\_PLAYER, REB\_PLAYER, TOV\_PLAYER: Aggregated player-level statistics per team.
* FG\_PCT, FG3M: Shooting efficiency indicators.
* PLUS\_MINUS: Captures the point differential when a team is on the court—an indirect measure of overall performance.

1. **Correlation Heatmap:**

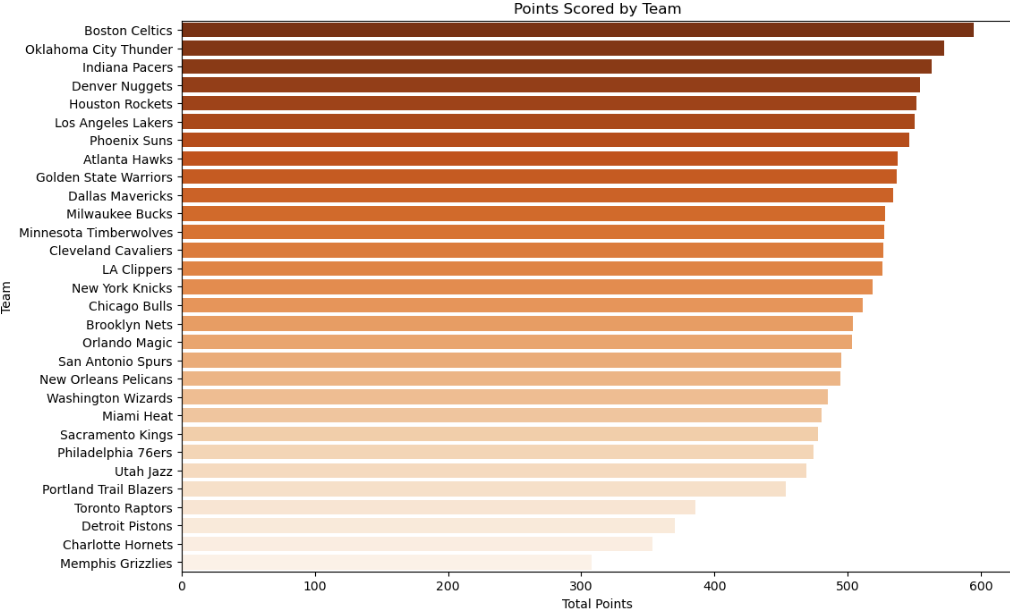
The heatmap revealed several strong correlations. Notably, Win Percentage was highly correlated with Points, Plus/Minus, and Playoff Rank, confirming that higher-scoring teams with better point differentials tend to perform better in the standings. Plus/Minus, in particular, emerged as one of the most predictive features of overall success, as it encapsulates both offensive and defensive performance. Turnovers showed a negative correlation with winning metrics, emphasizing the detrimental impact of poor ball control. This analysis supported our decision to include features like Points, Assists, Rebounds, Plus/Minus, and Turnovers in our modeling phase. The insights gained from the correlation heatmap played a critical role in validating our feature selection and enhancing the interpretability of our machine learning results.

1. **Bar Plot (Win Percent by Team):**

****The bar plot visualizes the win percentage of all 30 NBA teams for the 2023–2024 season. Teams are sorted in descending order, with darker shades indicating higher win rates. This visualization clearly distinguishes top-performing teams, such as the Boston Celtics, Denver Nuggets, and Oklahoma City Thunder, from lower-tier teams like the Detroit Pistons and Washington Wizards.

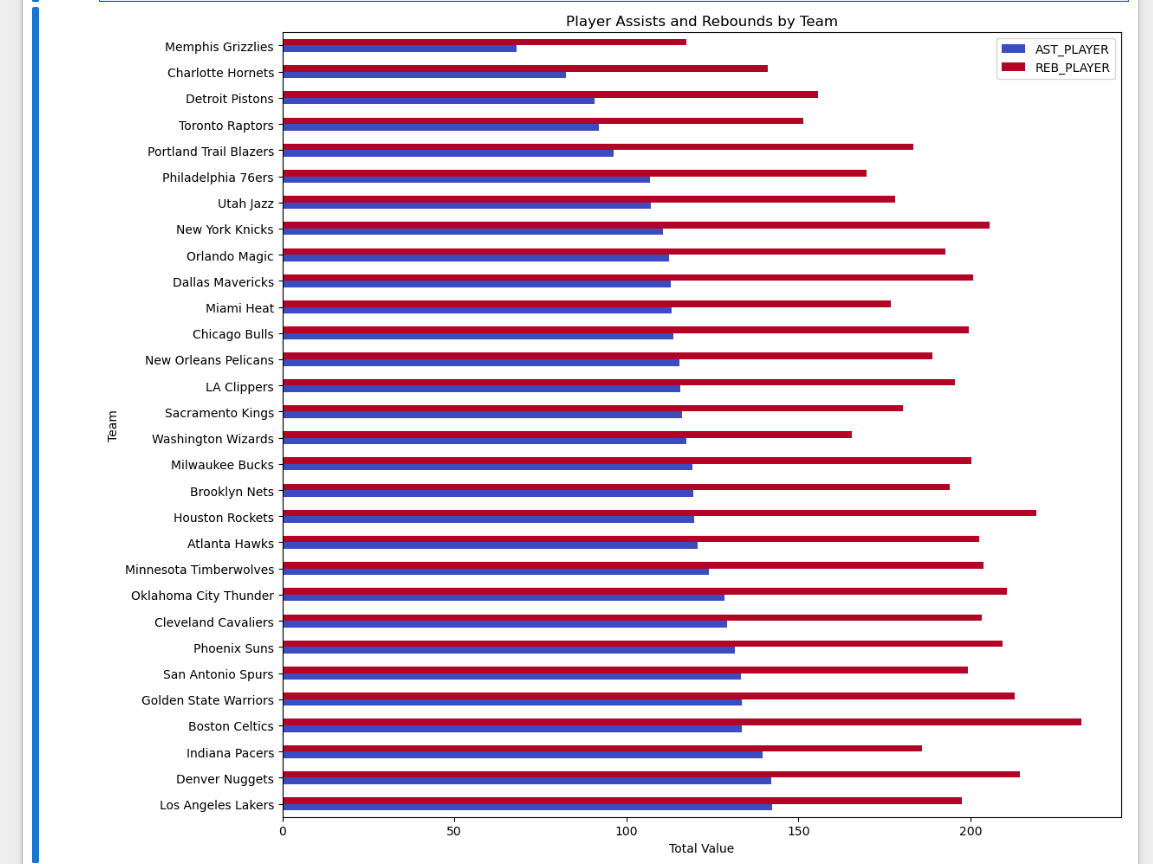
The chart offers an immediate understanding of the league’s competitive landscape and highlights the performance gap across teams. This distribution also informed our playoff group classification and served as a visual reference for model evaluation and clustering. Observing this spread reinforced the importance of win percentage as a key feature in predicting playoff success.

1. **Bar chart (Points Scored by Team):**

****This bar chart displays the total points scored by each NBA team, arranged in descending order. The visualization highlights offensive productivity across the league, with teams like the Boston Celtics, Oklahoma City Thunder, and Indiana Pacers leading in scoring output.

The gradient color scheme reinforces the scoring hierarchy, with darker bars representing higher totals. This plot supports the analytical findings that point production is strongly associated with win percentage and playoff success. The distribution also helps identify high-scoring teams that may underperform defensively, offering context for later modeling decisions.

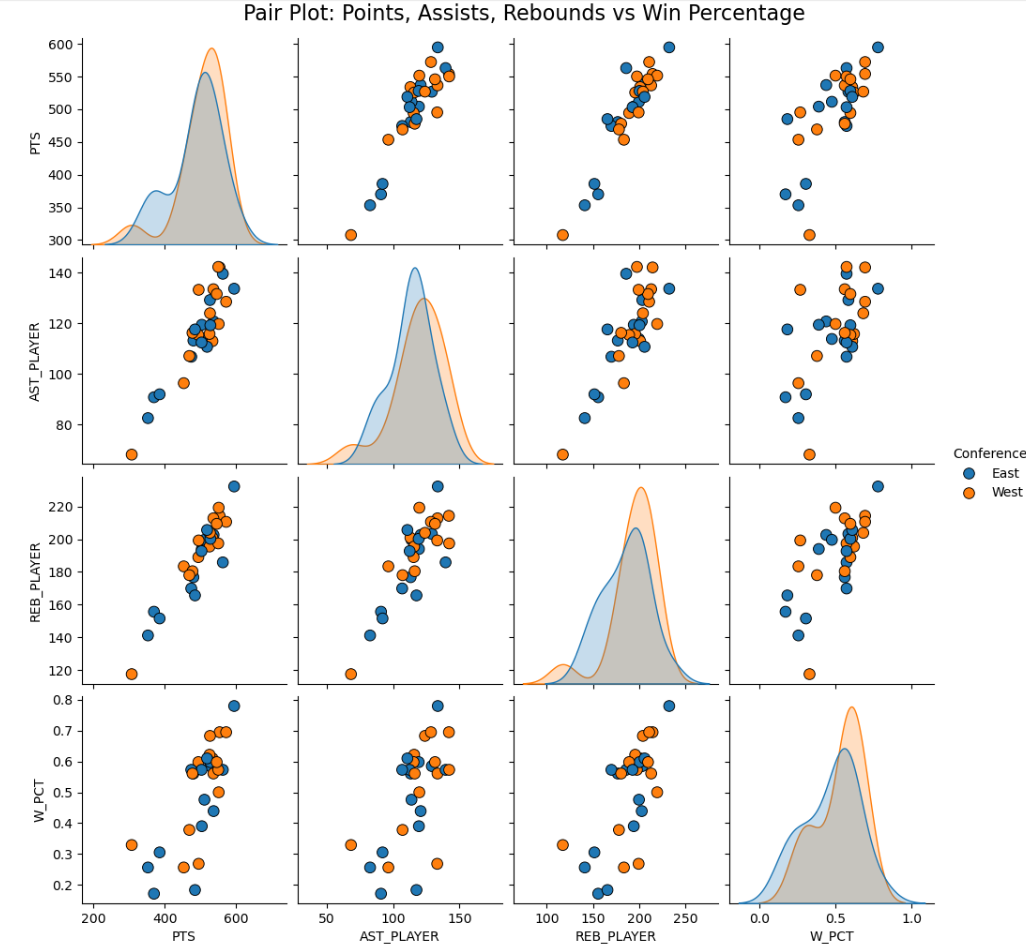
1. **Horizontal bar chart (Player Assists and Rebounds):**

****

This horizontal bar chart compares the total player assists and rebounds across all NBA teams. Assists are shown in blue, while rebounds are displayed in red, allowing for a side-by-side comparison of ball movement and board control. Teams like the Los Angeles Lakers, Denver Nuggets, and Indiana Pacers show strong values in both categories, indicating well-rounded performance.

The visualization helps identify teams with balanced offensive distribution and defensive presence—both key traits for playoff contention. Assists reflect team coordination and offensive strategy, while rebounds signal physical dominance and second-chance potential. These variables were included in the feature set due to their strong correlation with win percentage and playoff outcomes.

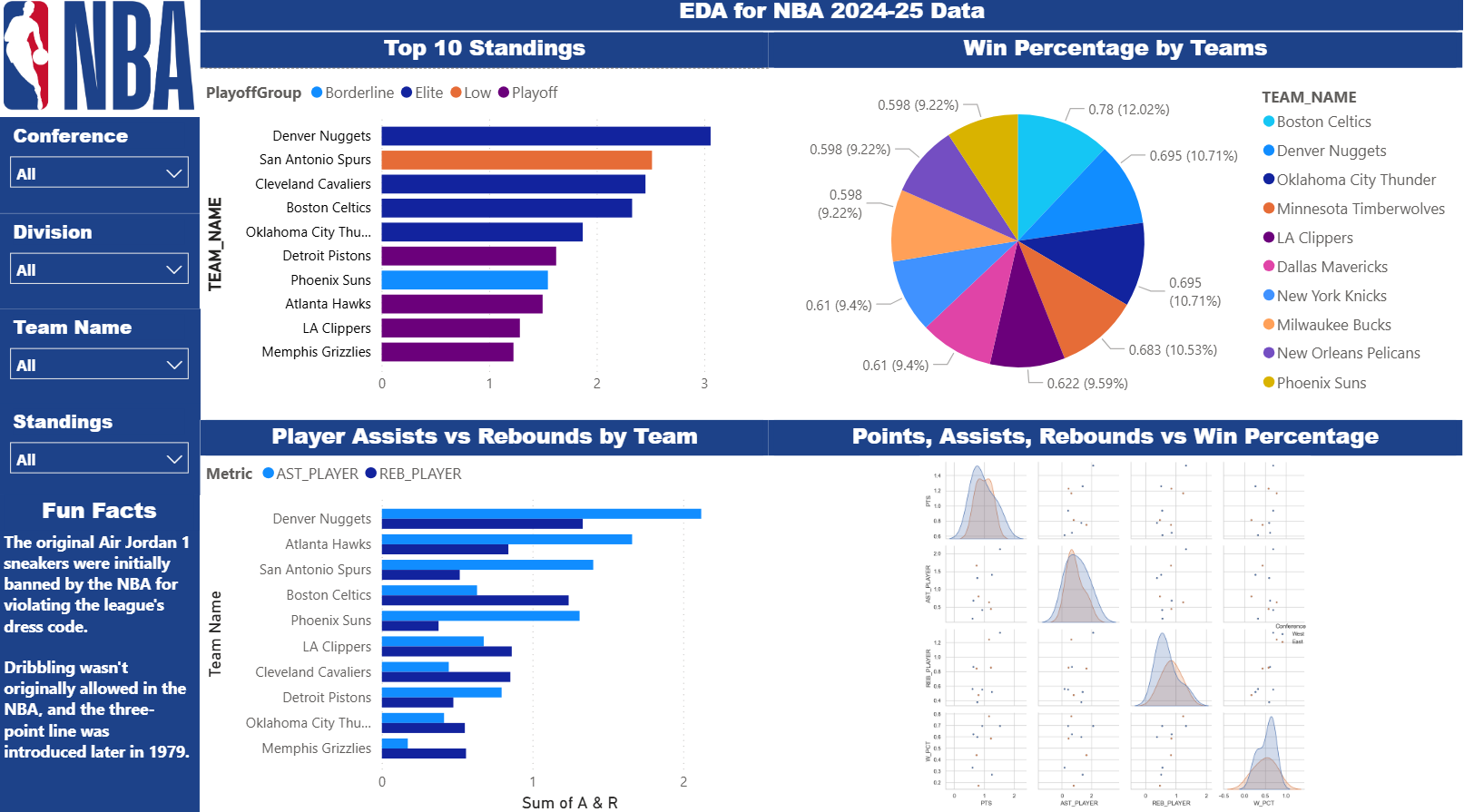
1. **Pair Plot (Points, Assists, Rebounds vs win percentage):**

****This pair plot visualizes the relationships between Points, Assists, Rebounds, and Win Percentage, with teams grouped by conference (East or West). The diagonal KDE plots show the distribution of each variable, while the scatter plots reveal their pairwise associations.

Overall, the plot demonstrates strong positive relationships among Points, Assists, and Win Percentage—teams that score more and pass effectively tend to win more games. Rebounds show a more moderate correlation, yet still contribute to winning trends. The separation between East and West teams is minimal, suggesting consistent patterns across conferences.

This visualization reinforces the importance of offensive efficiency and teamwork metrics in driving team success, supporting the features selected for clustering and classification models.

**Power BI EDA Dashboard overview:** [**Power BI Dashboard**](https://app.powerbi.com/groups/me/reports/eaa49883-aa29-4731-b2a0-7690b6968909/7d675c10dc0bb33c0784?ctid=5cdc5b43-d7be-4caa-8173-729e3b0a62d9&experience=power-bi&clientSideAuth=0)

****The interactive EDA dashboard provides a visual summary of key performance metrics for NBA teams during the 2024–2025 season. It is organized into four main sections:

* Top 10 Standings: Displays the top teams segmented by Playoff Group (Elite, Playoff, Borderline, Low), offering a quick view of predicted playoff tiers based on statistical clustering.
* Win Percentage by Teams: A pie chart highlighting the win rates of the top 10 teams. Each segment is color-coded to a team, offering a comparative look at win share distribution.
* Player Assists vs Rebounds by Team: A grouped bar chart comparing total assists and rebounds for the top 10 teams, helping visualize the balance between ball distribution and rebounding.
* Points, Assists, Rebounds vs Win Percentage: A compact pair plot analyzing correlations between major performance metrics and Win Percentage, with teams grouped by conference (East vs West).

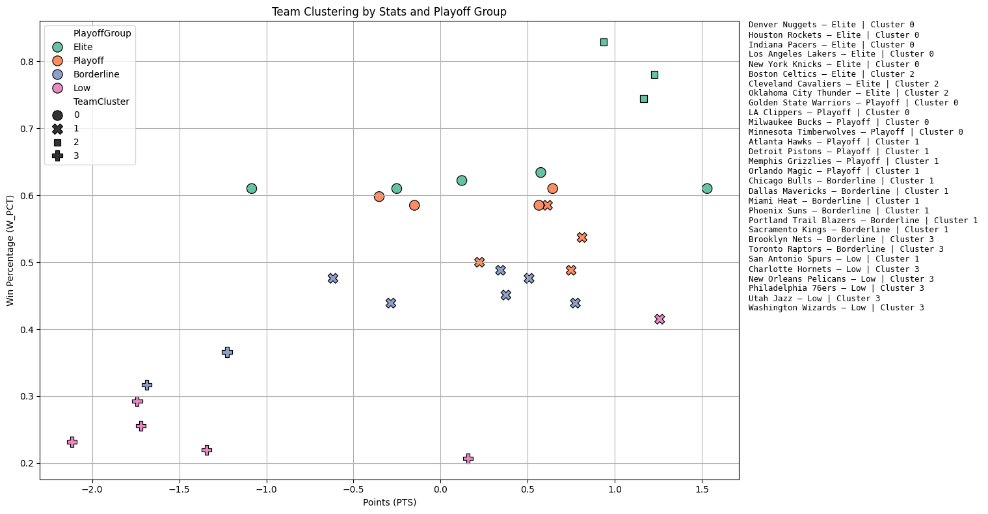
The dashboard included filtering options by conference, division, and team, as well as contextual NBA trivia. This layout enables both exploratory and comparative analysis, supporting pattern discovery and hypothesis validation for the predictive modeling phase.

1. **Feature Engineering**

To enhance model performance and provide more insightful inputs, two key features were engineered: Assist Ratio and Win Streak. Assist Ratio measures a team’s ability to generate assists without turning the ball over, serving as a proxy for team efficiency and smart decision-making. A high Assist Ratio reflects strong ball movement and fewer mistakes, characteristics often associated with successful teams. Win Streak, on the other hand, captures recent performance momentum. Teams on positive win streaks typically have a higher chance of success, especially approaching the playoffs. Including this feature helps the model detect trends in team dynamics beyond just season averages. These engineered metrics were critical in identifying performance patterns not captured by raw statistics alone.

1. **Segmentation Analysis (Clustering)**

To explore hidden patterns among teams, we applied K-Means Clustering to group NBA teams based on statistical similarity. Using scaled features such as points per game, win percentage, assist ratio, turnovers, steals, blocks, plus-minus, and win streak, the model identified four distinct clusters. These clusters represent performance tiers, which we manually aligned with playoff seed ranges: *Elite* (seeds 1–4), *Playoff* (5–8), *Borderline* (9–12), and *Low* (13–15). The Elbow Method was used to determine the optimal number of clusters (k=4). Feature values were standardized using StandardScaler to ensure equal weighting in distance calculations. This clustering step served both as exploratory analysis and as a way to validate our playoff group labeling for supervised model training.



The clustering scatter plot visualizes NBA teams based on their points per game (x-axis) and win percentage (y-axis), segmented into four clusters using K-Means. Each point represents a team, colored by their manually assigned playoff group (Elite, Playoff, Borderline, Low) and shaped by the cluster it belongs to. Teams in Cluster 0, like the Denver Nuggets and Lakers, exhibit high win rates and scoring, aligning with the "Elite" group. Cluster 1 contains solid playoff teams with decent performance, while Clusters 2 and 3 represent mid- to low-tier teams with lower win percentages and point totals. The plot supports the effectiveness of clustering in grouping teams by performance and confirms the relationship between statistical features and playoff positioning.

1. **Machine Learning Models:**

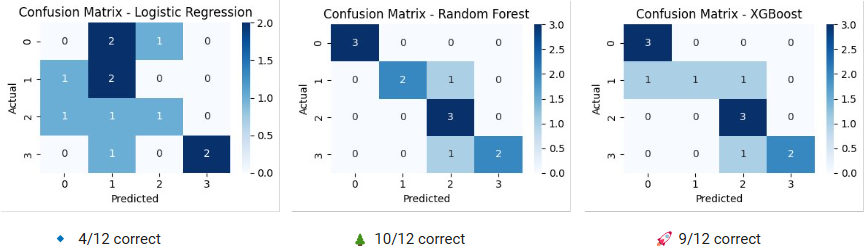
In the modeling phase, we implemented three classification algorithms to predict the NBA Playoff Group—a categorical target variable dividing teams into *Elite*, *Playoff*, *Borderline*, and *Low* tiers. These models include Logistic Regression, Random Forest, and XGBoost. The feature set used for training incorporated both raw statistics (e.g., *Points*, *Assists*, *Rebounds*, *Steals*, *Blocks*, *Turnovers*, *FG%*, *Plus/Minus*, *Win%*) and engineered features such as *Assist Ratio* and *Win Streak*, which offered deeper insights into team efficiency and momentum. To prepare for training, all features were normalized, and the dataset was split into a 60/40 stratified train-test split to maintain balanced class distributions.

**Performance Metrics:**

| **Model​** | **Accuracy​** | **Precision​** | **F1 Score​** |
| --- | --- | --- | --- |
| Logistic Regression​ | 0.416667​ | 0.458333​ | 0.411111​ |
| Random Forest​ | 0.833333​ | 0.900000​ | 0.837500​ |
| XGBoost​ | 0.750000​ | 0.837500​ | 0.726786 |

Logistic Regression was used as a baseline linear model but achieved only ~41.7% accuracy due to its inability to model complex relationships. Random Forest, a tree-based ensemble model, outperformed all others by effectively capturing non-linear patterns, reaching ~83.3% accuracy and an F1 score of 0.8375. Although XGBoost is typically a high-performing boosting algorithm, it delivered slightly lower results (~75% accuracy) in this case—possibly due to overfitting, lack of hyperparameter tuning, or the simplicity and size of the dataset. Ultimately, Random Forest proved to be the most robust and accurate model for classifying NBA teams into playoff tiers.

**Confusion Matrices**

****

The confusion matrices visually compare how well each model classified NBA teams into the four playoff tiers. Logistic Regression showed poor predictive power with only 4 out of 12 predictions correct, highlighting its struggle to capture complex team patterns. Random Forest outperformed all models, correctly predicting 10 out of 12 instances by effectively modeling nonlinear relationships between features. XGBoost also demonstrated strong performance with 9 out of 12 correct classifications, but slightly underperformed Random Forest. These results confirm that tree-based ensemble models are more suited for the multiclass classification of playoff groupings in this dataset.

**ROC Curve**

A graph of multi-class roc curve

AI-generated content may be incorrect.

The ROC curve offers a more nuanced evaluation of each model’s ability to separate classes. Random Forest achieved the best overall AUC scores across all classes, including a perfect AUC of 1.00 for Class 0 (Elite) and scores above 0.83 for other classes, demonstrating its strong and balanced classification ability. XGBoost followed closely with AUCs ranging from 0.81 to 0.93, indicating solid discriminative performance. Logistic Regression, however, performed inconsistently with AUC values as low as 0.48 for Class 1, reinforcing its limitations on this multiclass task. Overall, Random Forest provided the best trade-off between true positive rate and false positive rate across all playoff tiers.

1. **Predictions**

After evaluating model performance using key metrics such as Accuracy, Precision, F1 Score, and AUC, the Random Forest classifier emerged as the most effective model for predicting NBA playoff outcomes. It was subsequently chosen to generate the final predictions regarding playoff probabilities for the top-performing teams.

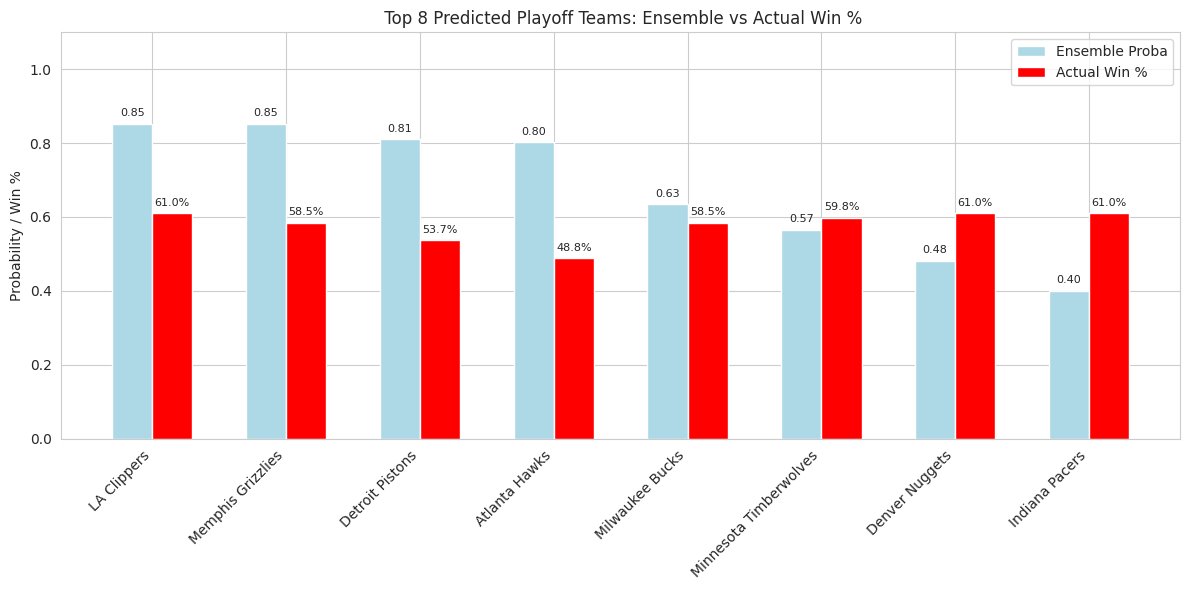
The final prediction chart highlights the top 8 NBA teams with the highest likelihood of making the playoffs, as predicted by the ensemble model. In this visualization:

* **Light blue bars** represent the **model’s predicted playoff probability**.
* **Red bars** denote the **actual win percentage** observed for each team.

This comparison allows us to assess how closely our model’s predictions align with real-world performance. For instance, the **LA Clippers** were predicted to have an **85% chance** of making the playoffs, while their actual win percentage stood at approximately **61%**. Despite the discrepancy in win percentage, the model effectively positioned them among the top-tier teams.

Similarly, the **Memphis Grizzlies** had a predicted playoff probability of **85%**, with an actual win rate of around **58.5%**. These examples illustrate the model’s ability to reasonably forecast playoff contenders even when exact win percentages vary. The consistency of ranking indicates that the model successfully captures underlying performance trends and dynamics.

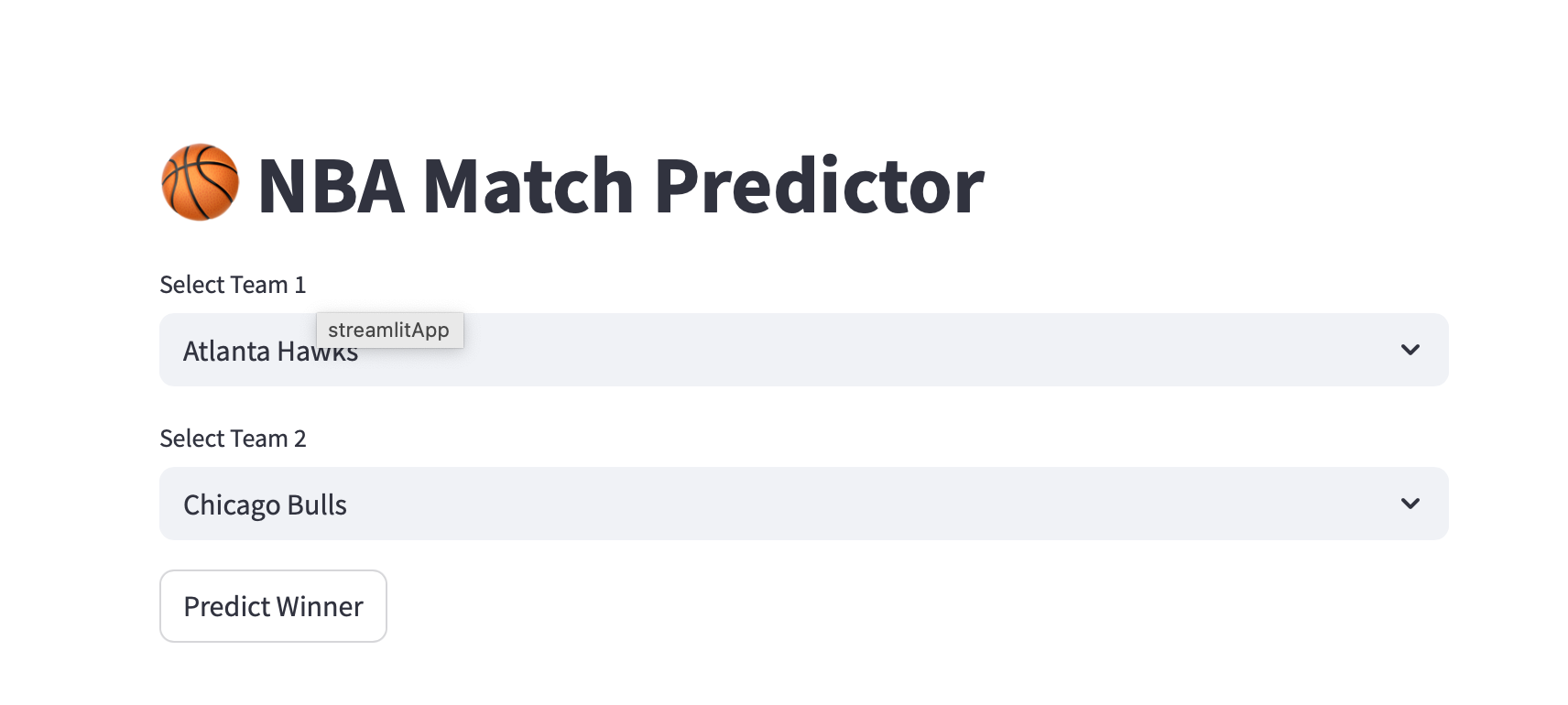
Overall, the ensemble model proved reliable in identifying the top playoff candidates, confirming its utility for predictive sports analytics.

****

**Team vs Team Predictions**

In addition to predicting overall playoff probabilities, we developed a **Team vs Team Matchup Predictor** that estimates the outcome of head-to-head NBA games using key performance metrics. This model was built using Random Forest and deployed via a **Streamlit-based web application**.

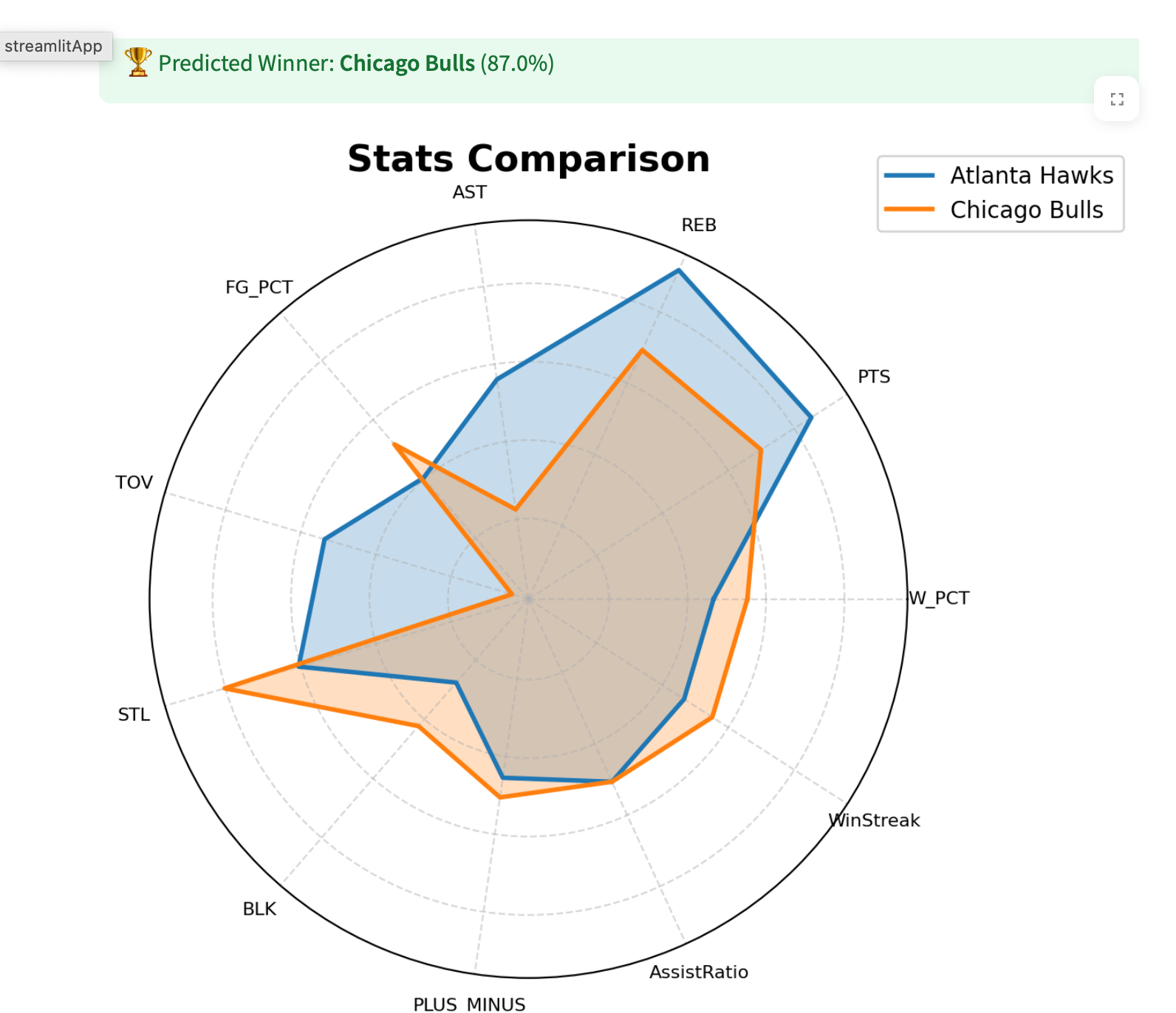
#### **Model Inputs:**

* Win Percentage
* Points Scored
* Rebounds
* Assists
* Turnovers
* Field Goal Percentage
* 

Users can select two teams from a dropdown menu within the web app. Upon selection, the application returns:

* The **predicted winning team**
* The **winning probability**
* Comparative metrics between both teams

**Example Output**:



* *Predicted Winner:* *Chicago Bulls*
* *Winning Probability:* 87%

This module serves as a practical extension of our project by enabling real-time matchup simulations and broadening the application of our models for interactive use.

Link to web app: [NBA Matchup Predictor](https://qhx2bpc6ewm23wnzjweqzb.streamlit.app)

1. **Conclusion**

The results of our modeling efforts demonstrate that machine learning can serve as a powerful tool for NBA championship and playoff prediction. Among the three models tested:- Logistic Regression, Random Forest, and XGBoost, tree-based ensemble methods (Random Forest and XGBoost) significantly outperformed the linear baseline model in terms of predictive performance.

The Random Forest classifier achieved the highest accuracy (83.3%) and F1 Score (0.8375), closely followed by XGBoost, confirming the superior capability of these models to capture complex, non-linear interactions inherent in basketball team data. These models demonstrated greater robustness and reliability, particularly in multi-class classification tasks involving playoff tier predictions.

A feature importance analysis further highlighted the key factors contributing to a team’s success:

* Points Scored
* Turnovers
* Win Percentage
* Rebounds and Assists

These features align well with basketball fundamentals, reinforcing the validity of the model's insights.

In contrast, Logistic Regression was limited by its linear nature and struggled to generalize in a setting characterized by intricate dependencies between performance metrics. Its lower accuracy (41.7%) and inconsistent classification results underscore its inadequacy for this problem space.

Overall, the success of the tree-based models supports the applicability of machine learning in sports analytics and paves the way for more sophisticated predictive systems in future iterations of this project.

1. **Future Work**

While the current model provides meaningful predictions and high accuracy, several avenues remain for improving and expanding this project:

1. **Hyperparameter Tuning & Cross-Validation:**  
   Implementing grid or randomized search to fine-tune model parameters and applying k-fold cross-validation would enhance model robustness and generalizability.
2. **Incorporation of Advanced Player Metrics:**  
   Future iterations could integrate metrics such as injury status, player synergy scores, bench depth, and minutes played, which significantly impact team performance but were not captured in the current model.
3. **Time-Series Modeling:**  
   Applying time-dependent models such as LSTM or ARIMA can help forecast how team performance evolves throughout the season, adding temporal insight to predictions.
4. **Interactive Dashboards:**  
   Developing a live, continuously updating dashboard with dynamic visualizations, real-time standings, and playoff simulations would make the tool more engaging and useful for both fans and analysts.
5. **Multi-Season Expansion:**  
   Training on a multi-season dataset can provide deeper generalization and better capture the variability in team and player dynamics across years.