NBA Player Stats & Team Win Prediction

Ryan Dielhenn, Momoka Aung, Angel Trujillo,
Jesus Villa, Harshil Patel

Introduction

Goals of the Project

- Understand how player and team performance indicators translate into scoring and winning
- Build interpretable models suitable for coaching, scouting, and predictive analysis

Why These Tasks Matter

- Player scoring helps with fantasy sports, rotations, and performance forecasting
- Team win prediction supports game strategy, lineup decisions, and opponent scouting

Contributions

- End-to-end pipeline: data cleaning → feature engineering → modeling → evaluation
- Benchmark models for both regression and classification tasks
- Insights showing which statistics influence performance most

Data Overview

Dataset: NBA 2024–2025 player game logs (NBA player stats dataset)

- **16,512 rows**, **25 features**
- Includes:
  - • Shooting stats (FG, FGA, 3P, 3PA, FT, FTA)
  - • Rebounds, assists, steals, blocks, turnovers
  - • Minutes played, team, opponent, and game result

Cleaning & Preparation

- Removed duplicates and non-play entries
- Engineered shooting accuracy stats (FG%, 3P%, FT%)
- No missing values after final preprocessing

2. Load and Explore Data

In [2]:
```python
# Load dataset (only need to do this once!)
player_stats_df = data_utils.load_nba_data()
```

```
Downloading dataset...
Path to dataset files: /Users/ryan/.cache/kagglehub/datasets/eduardopalmieri/nba-player-stats-season-2425/version
s/37

Available CSV files: ['database_24_25.csv']

============================================================================
DATASET OVERVIEW
============================================================================

Dataset shape: (16512, 25)

Column names:
['Player', 'Tm', 'Opp', 'Res', 'MP', 'FG', 'FGA', 'FG%', '3P', '3PA', '3P%', 'FT', 'FTA', 'FT%', 'ORB', 'DRB', 'TR
B', 'AST', 'STL', 'BLK', 'TOV', 'PF', 'PTS', 'GmSc', 'Data']

First few rows:
           Player   Tm  Opp Res     MP  FG  FGA    FG%  3P  3PA  ...  DRB  TRB  \
0   Jayson Tatum  BOS  NYK   W  30.30  14   18  0.778   8   11  ...    4    4
1  Anthony Davis  LAL  MIN   W  37.58  11   23  0.478   1    3  ...   13   16
2  Derrick White  BOS  NYK   W  26.63   8   13  0.615   6   10  ...    3    3
3   Jrue Holiday  BOS  NYK   W  30.52   7    9  0.778   4    6  ...    2    4
4  Miles McBride  NYK  BOS   L  25.85   8   10  0.800   4    5  ...    0    0

   AST  STL  BLK  TOV  PF  PTS  GmSc        Data
0   10    1    1    1   1   37  38.1  2024-10-22
1    4    1    3    1   1   36  34.0  2024-10-22
2    4    1    0    0   1   24  22.4  2024-10-22
3    4    1    0    0   2   18  19.5  2024-10-22
4    2    0    0    1   1   22  17.8  2024-10-22

[5 rows x 25 columns]

No missing values found!
```

Methodology

Player PTS Prediction (Regression)

- Target: Points scored (PTS)
- Models tested:
  Linear Regression
  Random Forest
  Gradient Boosting
  XGBoost
  LightGBM
- Train/test split: **60/40**

Team Win Prediction (Classification)

- Aggregated player data → team-level game metrics
- Target: Win (1) or Loss (0)
- Models tested:
  Logistic Regression
  Random Forest
  Gradient Boosting
  XGBoost
  LightGBM
- Balanced dataset among wins/losses

Evaluation Metrics

- **Regression:** RMSE, MAE, R^2
- **Classification:** Accuracy, Precision, Recall, F1, ROC-AUC

Feature Insights

Player-Level Findings

- **FGA (field goal attempts)** → strongest driver of scoring
- **3P made** and **minutes played** → major secondary predictors
- Efficiency metrics (3P%, FT%) help refine predictions
- Turnovers have a small negative impact on scoring

Team-Level Findings

- Higher **FG%**, **rebounds**, and **steals** correlate with higher win probability
- **Turnovers** consistently reduce win chances
- Teams winning the “efficiency battle” typically win the actual game

PTS Prediction Results

Performance Summary

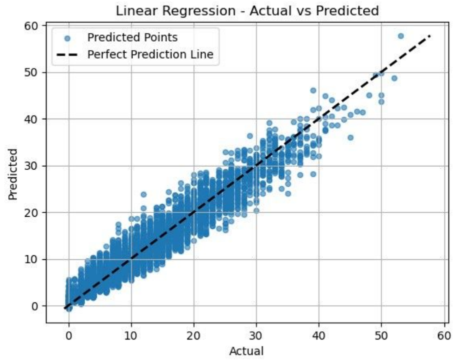
- Linear Regression: **RMSE = 2.17**, **MAE = 1.59**, **R^2 = 0.94**
- Gradient Boosting and Random Forest performed slightly worse

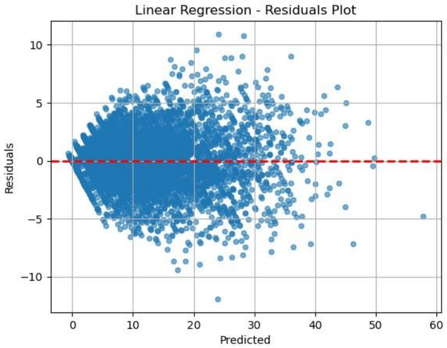
Interpretation

- Player scoring follows **mostly linear patterns**
- Scoring is dominated by **volume statistics** rather than complex interactions
- Nonlinear models capture some interactions but do not significantly improve performance

Key Takeaway

→ Player scoring is predictable using simple, interpretable features (shot attempts & minutes).





Win Prediction Results

Model Performance
Logistic Regression:

- Accuracy = **84%**
- Precision = **85%**
- ROC-AUC = **0.92**

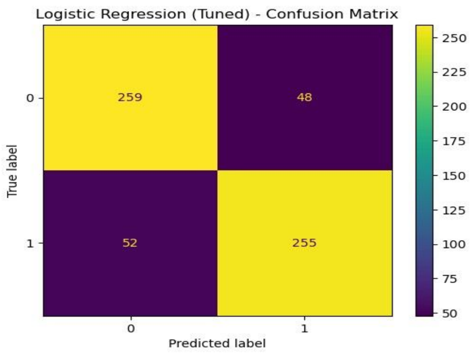
Random Forest & Gradient Boosting:

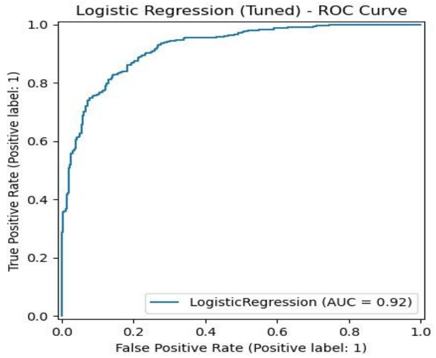
- Slightly lower accuracy (79–81%)
- Slightly lower ROC-AUC (0.87–0.89)

Interpretation

- Team victories depend heavily on **linear, interpretable factors**
- Shooting efficiency and ball control dominate prediction
- Complex models add nuance but not enough to outperform Logistic Regression

Key Takeaway
→ Simple models capture the majority of the win-loss signal.





Model Interpretation

PTS Model (Linear Regression)

- Strongest positive weights:
  • **FGA**, **3P**, **FT**
- Slight negative weight: **Turnovers**
- Confirms scoring is tied to shot volume and shooting quality

Win Model (Logistic Regression)

- Positive contributors:
  • **FG%**, **total rebounds**, **steals**
- Negative contributors:
  • **Turnovers**, **missed shots**
- Aligns with coaching principles: better shooting + extra possessions → wins

Model Evaluation

Hyperparameter Tuning and Cross Validation

- Added 5-fold cross-validation for all baseline and tuned models
- **Regression** models used **k-fold** validation while **classification** models used **StratifiedkFold** to maintain balance
- Applied **RandomizedSearchCV** for hyperparameter tuning using 10 iterations
- Ensured fair comparison and prevented misleading results

Key Findings (Regression - PTS Prediction)

- All tuned models improved over their baseline
- Biggest gain: **XGBoost** model(its default parameters may have been suboptimal for this task)
- RF, GB,LGBM showed smaller gains
- **Linear Regression** has minimal hyperparameters compared to other models and already achieved the best performance, so we did not attempt to tune it.

Key Findings (Classification - Win/Loss)

- Tuning provided minimal improvement
- **Baseline Logistic Regression** performed the best
- Some tuned models showed signs of overfitting
- The win/loss decision boundary appears to be linear, reducing the need for complex model configuration

Conclusion

Summary of Findings

- Built a complete NBA analytics pipeline
- Linear Regression best for predicting PTS
- Logistic Regression best for predicting wins
- Key performance indicators identified:
  • Shot volume & minutes → scoring
  • Shooting efficiency & turnovers → wins

Future Improvements

- Hyperparameter tuning (GridSearchCV, Bayesian search)
- Implementation of **XGBoost**, **LightGBM**, **CatBoost**
- Incorporate advanced stats (Player Impact Estimate, lineup data, pace)
- Add visualizations such as shot charts and correlation heatmaps