NBA Player Performance and Game Outcome Prediction - Progress Report

Team Members: Ryan Dielhenn, Momoka Aung, Angel Trujillo, Jesus Villa, Harshil Patel **Project Lead:** Ryan Dielhenn
Date: November 21, 2025

Project Overview

NBA Player Performance and Game Outcome Prediction aims to develop machine learning models that predict individual player scoring performance and team game outcomes using 2024-25 NBA season data. Our project employs two prediction tasks: (1) individual player scoring prediction (PTS) using regression models, and (2) team win/loss classification based on aggregated team statistics. By analyzing shooting efficiency, field goal attempts, assists, rebounds, turnovers, and playing time, we identify key factors that drive both individual scoring success and team victories, providing actionable insights for sports analytics and strategy optimization.

Team Member Responsibilities

- **Ryan Dielhenn (Project Lead):** Feature engineering, team data aggregation, classification models
- **Momoka Aung:** Documentation, feature engineering support
- **Angel Trujillo:** Exploratory data analysis, data visualizations
- **Jesus Villa:** Hyperparameter tuning, additional model implementations
- **Harshil Patel:** Performance analysis visualizations, metrics evaluation

---

Data Information

Source: Kaggle - NBA Player Stats Season 2024-25
Format: CSV (16,512 player-game records, 25 features)
Key Features: Scoring metrics (PTS, FG, FGA, 3P, FT), efficiency percentages (FG%, 3P%, FT%), playmaking (AST), defense (STL, BLK, TRB), usage metrics (MP, TOV, PF), and game context (team, opponent, result, date)
Target Variables: PTS (individual player scoring - regression), Team Win (game outcome - binary classification)
Data Quality: Complete dataset with no missing values, verified to contain 10-11 players per team per game enabling robust team-level aggregation

---

Project Status and Progress

We have established a complete data preprocessing pipeline and refactored our codebase into reusable functions that eliminate code duplication. Baseline regression models have been trained and evaluated for player scoring prediction (PTS). Linear Regression achieved the best performance ($R^2 = 0.939$, RMSE = 2.17, MAE = 1.59), with Field Goal Attempts (FGA) identified as the dominant predictive feature accounting for 78-79% of model importance, followed by 3-Point shooting percentage and Free Throw statistics. Gradient Boosting ($R^2 = 0.933$) and Random Forest ($R^2 = 0.927$) provided strong alternative models. Data structure analysis confirms the dataset contains 1,534 team-game combinations across 767 unique games, validating our approach for team-level win/loss classification. Our next phase includes exploratory data analysis with distribution and correlation analysis for both player-level and team-level data, feature engineering with categorical encoding and derived metrics, team-level data aggregation for binary classification, hyperparameter tuning via cross-validation for both regression and classification models, implementations of additional models, and comprehensive visualization of model performance and feature importance. The project foundation is complete with strong baseline results, positioning us well for these next steps.