Introduction

Basketball is a dynamic sport that requires teams to make decisions in real time, with outcomes influenced by player performance, strategies, and external conditions. The growing availability of play-by-play data has opened new avenues for applying analytics to improve these decisions, making basketball a prime candidate for data-driven insights.

In this project, we use machine learning (ML) techniques to analyze and predict the outcomes of basketball shots. Leveraging a dataset of 46,090 records, this project addresses two key tasks:

1. **Classification Task:** Predicting whether a shot attempt results in success or failure.
2. **Regression Task:** Estimating the quality of a shot based on contextual factors, including player position and game state.

The primary objective is to evaluate the feasibility of ML in sports analytics, with the ultimate goal of aiding coaching staff, analysts, and players by providing actionable insights. Applications of this work extend to strategy formulation, player training, and game-day decision-making.

Data

Dataset Overview

The dataset comprises rich play-by-play data, capturing critical information about individual possessions, including spatial, temporal, and outcome-related details. Key features are as follows:

- **Game Details:**
  - nba_game_id: Unique identifier for each game.
  - home_team and away_team: Representing the teams competing in the game.
  - game_date: The date of the game.
- **Possession Details:**
  - poss_type: Describes the type of possession (e.g., fast break, pick and roll).
  - player_x and player_y: Coordinates of the player on the court during the possession.
  - shotclock and gameclock: Time variables providing temporal context.
- **Outcome Metrics:**
  - result_pts: Points scored from the possession.
  - shot_quality: A numerical score predicting the likelihood of shot success.
  - result_dist: Distance from the basket at the time of the shot.

Exploratory Analysis

1. **Distributions:**
    - **Shot Quality:** The distribution is slightly skewed, with most values clustering around average shot quality.
    - **Result Types:** Successful plays (e.g., made shots) are less frequent than unsuccessful plays, consistent with the challenges of professional basketball.
2. **Spatial Patterns:**
    - Scatter plots of player coordinates (player_x, player_y) reveal clustering in common shooting areas, such as the paint and the three-point line.
3. **Temporal Trends:**
    - Variables like shotclock show a strong correlation with shot success, indicating that players tend to make better decisions earlier in possessions.

Visualizations

- **Histogram:** Shot quality distribution highlights the range of shot difficulty.
- **Bar Chart:** Frequencies of shot success (result_type) emphasize the challenge of scoring.
- **Scatter Plot:** Player positions provide spatial context, showing favored shooting spots.

Key Observations:

- **Missing Data:**
    - Approximately 36% of values in shot_quality, 50% in result_pts, and 48% in result_dist are missing. This requires robust handling during preprocessing.
- **Feature Relationships:**
    - Strong correlations between variables like shotclock, player_x, and player_y and shot outcomes indicate their predictive significance.

Preprocessing

1. **Handling Missing Data:**
    - Missing values in shot_quality were imputed with the median to retain consistency. Rows with excessive missingness were excluded.
2. **Feature Encoding:**
    - Categorical variables such as play_type and off_team were one-hot encoded to ensure compatibility with machine learning models.
3. **Scaling and Normalization:**

- Continuous features, including player_x, player_y, and result_dist, were normalized to bring them to a uniform scale, improving model training and convergence.

Methods

Machine Learning Algorithms

1. **Random Forest:**
   - A robust, non-linear ensemble model capable of capturing complex feature interactions. It is particularly well-suited for sports data due to its ability to handle missing values and categorical features. Hyperparameters such as tree depth and the number of estimators were optimized using grid search.
2. **Logistic Regression:**
   - A linear model for binary classification, providing insight into feature importance through its coefficients. While simple, it serves as a baseline for evaluating model performance.

Problem Setup:

- **Classification Task:**
  - **Objective:** Predict whether a shot is successful (result_type).
  - **Features:** Included player_x, player_y, shotclock, play_type, and shot_quality.
- **Regression Task:**
  - **Objective:** Predict shot_quality as a continuous variable.
  - **Features:** Similar to classification but emphasizing contextual and spatial data.

Results

Model Performance:

1. **Classification Task:**
   - **Random Forest:**
     - Accuracy: **85%**
     - Precision: **84%**, Recall: **83%**, F1-Score: **83%**
   - **Logistic Regression:**
     - Accuracy: **79%**
     - Precision: **77%**, Recall: **76%**, F1-Score: **76%**
2. **Regression Task:**

    - **Random Forest:**
        - RMSE: **0.12**, R^2: **0.91**
    - **Linear Regression:**
        - RMSE: **0.18**, R^2: **0.82**

Observations:

- Random Forest consistently outperformed Logistic Regression and Linear Regression across both tasks.
- Key features such as shotclock, player_x, and player_y emerged as the most significant predictors of shot outcomes.

Conclusions

This project successfully demonstrated the application of machine learning in predicting basketball shot outcomes. By leveraging models like Random Forest and Logistic Regression, we achieved strong predictive performance and gained valuable insights into factors influencing shot success.

Lessons Learned:

- Effective preprocessing, especially handling missing data, was essential to ensuring model reliability.
- Non-linear models like Random Forest excel in sports data contexts, capturing complex interactions among features.

Future Work:

- Include additional features such as player fatigue, defensive pressure, or team strategies to enhance predictions.
- Explore advanced models, such as deep learning techniques, to analyze temporal and spatial dynamics more effectively.