Methodology Overview

Introduction

- Overview: predicting each players' three-point % based on the file fas_2024.csv
- Dataset: 108 players

Model Choice

- **Multiple linear regression model**: when dealing with a small dataset, we should use a simple model to avoid overfitting

Data Exploration

- **Correlation heatmap:**

Correlation between features:

1. Corr(three_pct_season(y), three_cnr_pct_oct_nov(x)) = 0.508

As the player's field goal % from corner three-point shots increases, there's a tendency for the overall three-point % to increase

implies the importance of strategic shooting

2. Corr(lwr_paint_shots_oct_nov(x), ft_shots_oct_nov(x)) = 0.736

potential multicollinearity issue

- **Histograms** for all potential x variables:
  - reveals significant differences in scales
  - need for **standardization** (also for better interpretability)

Model Training Workflow

01 Standardization of features

I. Omitting non-numerical ‘Name’ column during data processing
II. combined_df for concating ‘three_pct_season’ + 12 standardized features

02 Splitting data into training & testing sets

I. Prepared for running regression models & later K fold validation

03 K Fold Validation

- **5-fold** validation chosen because of the dataset size
- **Cross-Validation MSE Function**:
  define cross_val_mse_test() for k-fold cross-validation to calculate average MSE
- **Feature selection loop**:
  iterated through all possible feature combinations to find the subset minimizing MSE
- **Result Output**:
  printed and reported the best feature combination and its associated minimum MSE

04 Linear Regression Model Results

I. Run the multiple linear regression model using for the best feature combinations

Prediction Accuracy & Results

Feature Selection:

Best feature combination =

(Coefficient) (Intercept / Variable)

0.361130 * Intercept +
0.006922 * upr_paint_pct_oct_nov +
0.005754 * mid_pct_oct_nov +
0.013536 * three_non_cnr_pct_oct_nov +
0.015440 * three_cnr_pct_oct_nov +
0.010103 * ft_pct_oct_nov +
-0.005966 * lwr_paint_shots_oct_nov +
-0.004134 * mid_shots_oct_nov +
0.007972 * three_non_cnr_shots_oct_nov

Performance Metric: Mean Square Error

MSE is suitable because it:

- quantifies prediction accuracy
- is sensitive to deviations
- robustly handles outliers

Minimum Cross-Validation MSE: 0.000886

- emphasizes the model's strong predictive capacity on unseen data

Inclusion of all variables	Best feature combination
MSE on training: 0.000714	MSE on training: 0.000757
MSE on validation: **0.000981**	MSE on validation: **0.000791**

Feature selection demonstrates enhanced predictive capacity, optimizing the model's performance.