

Basketball Performance Analysis Project

Master's Level Data Science Exercise

Project Overview

Objective

This project aims to analyze basketball team and player performance using a comprehensive synthetic dataset. You will apply advanced statistical modeling techniques to understand the factors influencing team success, player performance, and championship potential.

Dataset

A comprehensive synthetic dataset (`basketball_performance_dataset.csv`) is provided, capturing multidimensional aspects of basketball performance across teams, players, and seasons.

Data Dictionary

Basketball Performance Dataset (`basketball_performance_dataset.csv`)

Team and Organizational Variables:

- `season`: Year of the basketball season
- `team_id`: Unique identifier for each team
- `team_name`: Name of the basketball team
- `team_budget`: Annual team budget (USD)
- `market_size`: Team's market size (Small/Medium/Large)
- `conference`: League conference (Eastern/Western)
- `total_team_wins`: Number of wins in the season
- `playoff_appearance`: Binary indicator of playoff qualification

Player Demographic Variables:

- **player_id**: Unique identifier for each player
- **player_age**: Player's age
- **position**: Player's primary position (Guard/Forward/Center)
- **height_cm**: Player's height in centimeters
- **years_in_league**: Number of years playing professionally

Performance Metrics:

- **games_played**: Number of games played in the season
- **points_per_game**: Average points scored per game
- **assists_per_game**: Average assists per game
- **rebounds_per_game**: Average rebounds per game
- **performance_score**: Composite performance metric
- **player_salary**: Annual player salary (USD)

Career and Resilience Indicators:

- **career_trajectory**: Player's career progression (-1: declining, 0: stable, 1: emerging)
- **injury_count**: Number of injuries in the season
- **injury_impact**: Estimated impact of injuries on performance

Qualitative Data:

- **performance_narrative**: Descriptive text about player's performance characteristics

Prediction Target:

- **championship_potential**: Binary indicator of potential championship success

Analysis Tasks

Task 1: Exploratory Data Analysis (20%)

1. Comprehensive exploratory analysis
 - Descriptive statistics for performance metrics
 - Distribution analysis of team and player variables
 - Correlation matrix between performance indicators
2. Visualization Requirements:
 - Performance variations across positions
 - Salary and performance relationships
 - Team performance by conference and market size

Task 2: Regression Modeling (40%)

Objective: Develop regression models to predict player and team performance

1. Linear Regression: Predicting Performance Score

- Dependent Variable: `performance_score`
- Independent Variables:
 - `player_age`
 - `years_in_league`
 - `height_cm`
 - Performance metrics (points, assists, rebounds)
 - Team-level features

Requirements:

- Implement multiple linear regression
- Check and address multicollinearity
- Validate model assumptions
- Interpret coefficients and statistical significance

2. Regularized Regression

- Apply Ridge or Lasso regression
- Compare model performance metrics
- Discuss feature importance

Task 3: Binary Outcome Prediction (40%)

Objective: Predict championship potential

- Dependent Variable: `championship_potential`
- Independent Variables:
 - Individual player performance metrics
 - Team-level characteristics
 - Career trajectory
 - Injury-related features

Requirements:

- Implement both Logit and Probit models
- Compare model performance using:
 - Accuracy
 - AUC-ROC

- Confusion matrix
- Interpret marginal effects
- Discuss model selection criteria

Submission Requirements

1. Comprehensive analysis report (max 15 pages)
2. Fully documented code
3. Detailed result interpretations
4. Discussion of limitations and potential improvements

Evaluation Criteria

- Technical Complexity (40%)
- Statistical Rigor (30%)
- Visualization Quality (15%)
- Interpretation Depth (15%)

Bonus Challenges

1. Provide a theoretical framework for player performance prediction
2. Develop advanced feature engineering techniques
3. Explore non-linear relationships in performance prediction
4. Create predictive models for player career trajectories