Machine Learning Football Predictor:
English Premier League Match Outcome Prediction

Shinkre Ved Rahul

This project implements a machine learning system to predict match outcomes in the English Premier League. Combining web scraping techniques with supervised learning algorithms, the system analyzes historical match data to forecast home wins, away wins, or draws. The project demonstrates data science applications in sports analytics while addressing football's unique prediction challenges.

Introduction

Motivation

Football has long been a domain where statistical analysis and intuition intersect. As a lifelong football enthusiast and data scientist, I developed this predictive model to bridge these worlds. The English Premier League's competitiveness and data availability make it ideal for analysis.

Project Objectives

- Collect and process historical EPL match data
- Engineer relevant predictive features
- Implement and evaluate machine learning models
- Develop an accurate prediction system

Data Collection and Preparation

Data Source

Selected FBref for:

- Comprehensive historical coverage
- Consistent data formatting
- Advanced metrics availability
- Free educational access

Data Cleaning

Key preprocessing steps:

- Handling missing values
- Standardizing team names
- Creating derived features
- Normalizing numerical features

Feature Engineering

Predictor Sets

- **Baseline Predictors**: Venue, opponent, time, day
- **Rolling Averages**: Goals, shots, possession metrics
- **Full Feature Set**: Baseline + rolling + FIFA rankings

Target Variable

Match outcomes encoded as:

- 0: Away win
- 1: Draw
- 2: Home win

Model Selection

Candidate Models

- **Basic Models**: KNN, LDA, QDA, Logistic Regression
- **Tree-Based Models**: Decision Trees, Random Forest, Gradient Boosting

Evaluation Metrics

- Accuracy
- Precision
- Training vs testing performance

Results

Optimal Models

- **Primary Recommendation**: Random Forest (Full Feature Set)
  - Testing Accuracy: 65.93%

    - Testing Precision: 64.64%
    - Best generalization capability

- **Secondary Recommendation**: K-Nearest Neighbors (Rolling Features)
    - Training Accuracy: 74.17%
    - Training Precision: 76.39%
    - Best training performance

Key Findings

- Strong home advantage prediction (70% accuracy)
- Draws most challenging to predict (50%)
- Recent form more predictive than long-term history

Challenges and Future Work

Limitations

- Missing data in early seasons
- Football's inherent randomness
- Team strength fluctuations

Future Improvements

- Incorporate player-level data
- Add betting odds benchmarks
- Implement temporal pattern recognition

Conclusion

This project demonstrates machine learning's potential for football outcome prediction, with Random Forest emerging as the most reliable approach , yet future enhancements could always be done.