Premier League xG Prediction Model - Project Report

1. Introduction

This project aims to develop a predictive model for estimating the expected goals (xG) in English Premier League (EPL) matches. Expected Goals is a key metric used in football analytics to quantify the quality of scoring chances. By leveraging historical shot-level data, match outcomes, and rolling team metrics, this model provides a data-driven foundation to simulate future match outcomes and understand team performance trends.

2. Data Collection

The data for this project was sourced from Understat, a popular football analytics website. A custom Python web scraping script was created to collect shot-level data for the 2023/24 Premier League season. Each shot included key attributes such as xG, shot location, player, match context, and result.

3. Data Processing and Aggregation

The raw shot-level data was transformed into a match-level dataset. This involved aggregating shots by team and match to calculate:
- Total xG (sum of xG values per match)
- Shot count
- Conversion rate (goals divided by shots)
- Average xG per shot
- Total goals scored

Separate summaries were created for home and away performances using rolling metrics over the last 5 matches. These rolling metrics included offensive and defensive statistics such as rolling xG, rolling shots, goals conceded, and opponent conversion rates.

4. Feature Engineering

The rolling home and away metrics were merged based on match fixtures to build the final training dataset. This dataset included all relevant predictors for both teams including:
- Rolling offensive metrics: xG, shots, average shot xG, conversion rate
- Rolling defensive metrics: shots conceded, goals conceded, opponent conversion rate

The dataset was organized such that each row represented a match with the home and away team features clearly aligned.

5. Rationale for Methods Used:

For evaluating predictive regression models (like xG), the metrics you've used:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)

are indeed standard and widely used in both academia and industry. Here's how they fit into the industry context:

Metric	Use Case	Pros	Cons
MAE	Evaluating average absolute prediction error	Easy to interpret, robust to outliers	Treats all errors equally
RMSE	Penalizing large errors more	Sensitive to large mistakes, good for risk-averse scenarios	Can exaggerate impact of outliers

6. Model Development

Separate regression models were trained to predict the number of goals scored by the home and away teams using Gradient Boosting Regressor. Features used included rolling offensive and defensive metrics. The dataset was split into training and test sets to validate model performance.

7. Model Performance

The model was evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The following performance was observed on the test set:
- Home Goals Prediction MAE: 1.083
- Home Goals Prediction RMSE: 1.375
- Away Goals Prediction MAE: 1.011
- Away Goals Prediction RMSE: 1.324

8. Conclusion

This project demonstrates a comprehensive approach to building a football analytics model using xG. From web scraping and feature engineering to training predictive models, the pipeline allows us to evaluate teams based on historical performance and simulate match outcomes. The current model forms a strong foundation for future enhancements including player-level modeling, betting strategies, and match simulations.