

# English Premier League (EPL) scoreline prediction model



By Uyi Erhabor  
July 2025

# Introduction

## **Presentation structure:**

- Methodology
- Key Visualisation
- Model selection
- Model Evaluation

# Methodology

## **Datasets:**

- Match data: epl\_matches\_messy
- Team attributes data: team\_attributes\_messy
- The combined dataset covered the 2019/20 – 2023/24 seasons

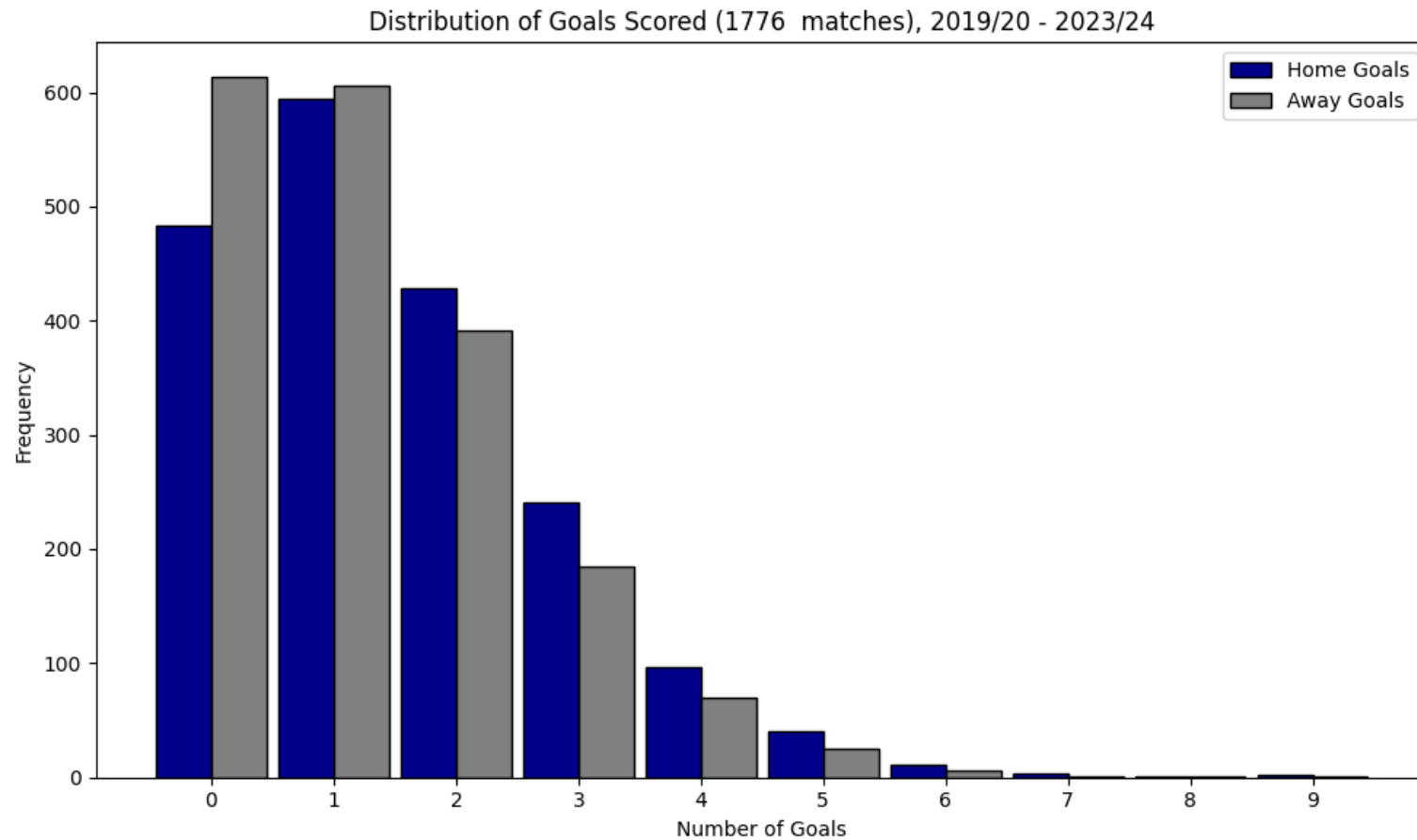
## **Data Preparation:**

- Carried out cleaning on the columns in each dataset
- Explored the dataset by deriving useful information on each column
- Created features from each dataset
- Combined match and team attributes data

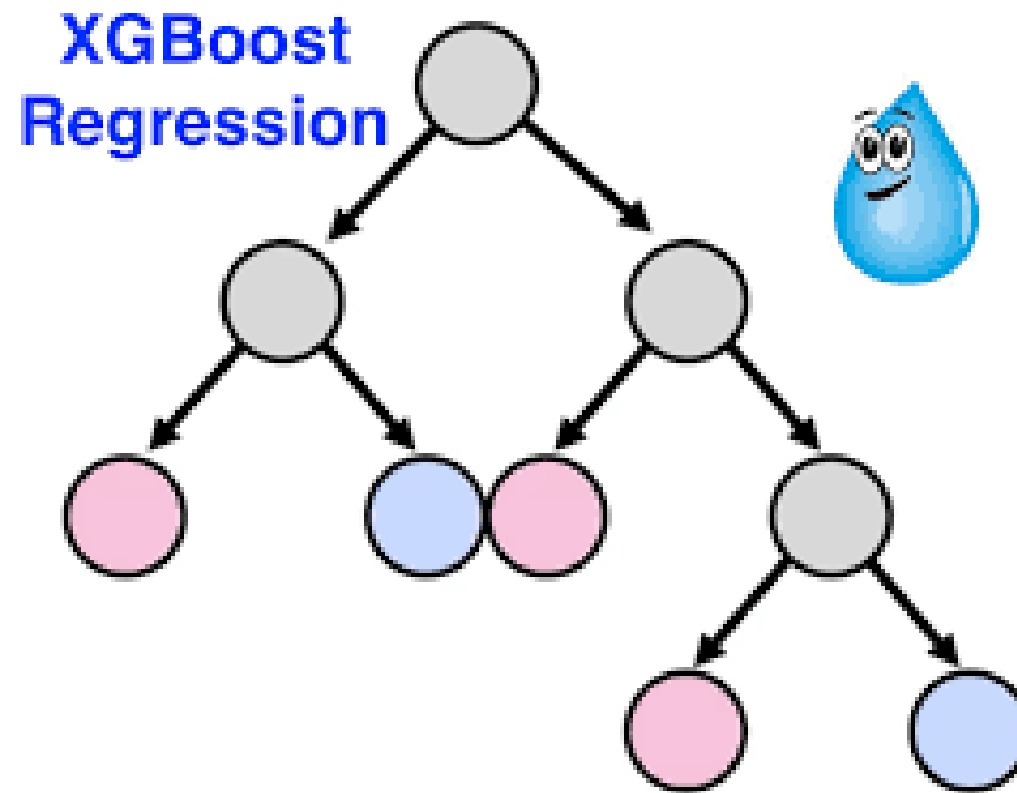
## **Analysis:**

- Built an XG Boost model to predict scorelines
- Evaluated model

# Distribution of goals



# Model selection



# Model Evaluation

**R-squared:** a measure of the combined importance of independent variables in a model in determining one or more dependent variables

**R-squared (Home/Away):** The R-squared value implies that the model provides a baseline for future predictions with additional features required to provide more robust scoreline predictions

**Mean Absolute Error:** a measure of the average absolute difference between predicted and actual values

**Mean Absolute Error (MAE) (Home/Away):** On average, the model's prediction of the number of goals the home team and away teams score is typically **+/- 1 goal** of the actual number of goals scored

**Overall Mean Absolute Error (MAE):** The total average prediction error for a match's scoreline is approximately **+/- 1 goal**