Integrating Team Statistics and Betting Odds for Premier League Match Predictions

Prasanna Shrestha

October 17, 2025

1 Introduction

In this study, I have used past match data and bet odd scores of major agencies to train a neural network ,logistic regression, and random forest model that predicts match outcomes — whether it'll be a home win, draw, or away win. The dataset includes features like match goals,shots attempted, and betting scores from agencies such as B365, 1xbet, spanning from 2000 to 2025. After training the model, predictions were done on upcoming fixtures like Manchester United vs Liverpool,Man City vs Everton, showing the predicted result along with the win probabilities for each outcome.

2 Dataset Description

The models were trained on English Premier League datasets spanning from season 2000-2001 to 2025-2026,having approximately 10000 matches. Each row correspond to match statistics with features such as:

- home and away teams
- Full time home and away goals (FTHG and FTAG),
- shots stats (HS,AS & HST,AST),
- corners(HC & AC)
- Home and Away fouls (HF & AF)
- red and yellow cards (HR & AR, HY &AY)
- referees
- bet odds score from major agencies i.eB365, BWD (avgH, avgD, avgA)
- Full time result (FTR)

3 Feature Engineering

Raw features including shot statistics, betting odds, and team form were transformed into more informative metrics to improve prediction performance.

3.1 Head-to-Head Performance Metrics

Win and draw rates for home (H) and away (A) teams are computed as:

$$\text{WinRate}_X = \frac{\text{Total Wins}_X}{\text{Number of Matches between }_X} \tag{1}$$

$$\text{DrawRate}_X = \frac{\text{Total Draws}_X}{\text{Number of Matches between}_X}, \quad X \in \{H, A\} \tag{2}$$

Comparative gap features capture the difference between home and away team performance:

$$\text{GoalDiffGap} = \text{HomeGoalDiff} - \text{AwayGoalDiff} \tag{3}$$

$$\text{PointsGap} = \text{HomePoints} - \text{AwayPoints} \tag{4}$$

$$\text{FormGap} = \text{HomeForm} - \text{AwayForm} \tag{5}$$

$$\text{WinRateGap} = \text{HomeWinRate}_{L5} - \text{AwayWinRate}_{L5} \tag{6}$$

$$\text{DrawRateGap} = \text{HomeDrawRate}_{L5} - \text{AwayDrawRate}_{L5} \tag{7}$$

3.2 Efficiency and Shot Statistics

Shot conversion efficiency measures how effectively a team converts its total shots into shots on target. It is defined as the ratio of shots on target to total shots for each team:

$$\text{ShotEfficiency}_{\text{Home}} = \frac{\text{HomeShotsOnTarget}}{\text{HomeTotalShots}} \tag{8}$$

$$\text{ShotEfficiency}_{\text{Away}} = \frac{\text{AwayShotsOnTarget}}{\text{AwayTotalShots}} \tag{9}$$

3.3 Betting Market Indicators

3.3.1 Implied Probabilities

Betting odds are converted to normalized implied probabilities using the following formula, where h, a, and d denote home win, away win, and draw odds respectively:

$$\text{ImpliedProb}_{\text{Home}} = \frac{1/h}{1/h + 1/a + 1/d} \tag{10}$$

$$\text{ImpliedProb}_{\text{Draw}} = \frac{1/d}{1/h + 1/a + 1/d} \tag{11}$$

$$\text{ImpliedProb}_{\text{Away}} = \frac{1/a}{1/h + 1/a + 1/d} \tag{12}$$

3.3.2 Odds-Derived Features

Market consensus and uncertainty are captured through odds ratios and variance metrics:

$$\text{HomeAwayWinRatio} = \frac{\text{AvgHome}}{\text{AvgAway}} \tag{13}$$

$$\text{DrawOddRatio} = \frac{\text{AvgDraw}}{\min(\text{AvgHome}, \text{AvgAway})} \tag{14}$$

$$\text{HomeOddVariance}, \quad \text{AwayOddVariance}, \quad \text{BettingConfidence} \tag{15}$$

$$\text{FavoriteStrength}, \quad \text{UnderdogStrength} \tag{16}$$

3.4 Scoring Pattern Analysis

Low-scoring match tendencies for each team are tracked along with comparative gaps:

$$\text{HomeLowScoreRate}, \quad \text{AwayLowScoreRate} \tag{17}$$

$$\text{LowScoreGap} = \text{HomeLowScoreRate} - \text{AwayLowScoreRate} \tag{18}$$

4 Final Features

- Head-to-head win rates:
  - `home_vs_away_winrate`
  - `away_vs_home_winrate`
- Match performance stats:
  - `home_shot_efficiency`
  - `away_shot_efficiency`
- Last 5 match stats:
  - `home_draw_rate_last5`
  - `away_draw_rate_last5`
  - `draw_rate_gap`
  - `home_low_score_rate`
  - `away_low_score_rate`
  - `low_score_gap`
- Gap features:
  - `goal_diff_gap`
  - `points_gap`
  - `form_gap`
  - `win_rate_gap`
  - `attack_strength_gap`
  - `defense_strength_gap`
  - `league_position_gap`
- Betting odds:
  - `home_away_win_ratio`
  - `draw_odd_ratio`
  - `h_implied_prob`
  - `a_implied_prob`
  - `d_implied_prob`
  - `home_odd_variance`
  - `away_odd_variance`
  - `betting_confidence`
  - `favorite_strength`
  - `underdog_strength`
- Target:
  - `FTR` (Full Time Result)

5 Modeling Approach

Two machine learning models were trained for match outcome prediction.Both models were trained on historical data with features including team form, head-to-head stats, match averages, and betting odds.

5.1 Neural Network

- Neural network with 3 hidden layers (128, 64 and 32 neurons) with batch dropout of 0.3
- Activation function: ReLU for hidden layers, Softmax for output
- Optimizer: Adam
- Loss function: Sparse Categorical Cross-Entropy
- Metrics: Accuracy
- Batch size: 32, Epochs: 60
- Class weights were used to handle unbalanced target classes, especially for the underrepresented **"Draw"** class

5.2 Random Forest Classifier

- Numbers of estimators: 300
- Max Tree depth: 12
- Min sample split: 5
- Min sample leaf: 2
- Class weight: Balanced

5.3 Logistic Regression

- Solver: `lbfgs`
- Regularization: L2 penalty to prevent overfitting
- Class weights: Balanced to handle unbalanced target classes, particularly the underrepresented **"Draw"** class
- Metrics: Accuracy

6 Data preparation

- The dataset was split into training (80%) and test (20%) sets.
- Numeric features containing infinite or NaN values were replaced with their respective means. Furthermore, a **Standard Scaler** was used to scale these values.
- The target labels of the FTR were encoded as: $H \to 0,\ D \to 1,\ A \to 2$.

7 Results

7.1 Classification Report

Table 1: Random Forest

Class	Precision	Recall	F1-score	Support
0	0.67	0.71	0.69	869
1	0.31	0.24	0.27	467
2	0.52	0.56	0.54	560
Accuracy		0.55		1896
Macro Avg	0.50	0.51	0.50	1896
Weighted Avg	0.54	0.55	0.54	1896

Table 2: Neural Network

Class	Precision	Recall	F1-score	Support
0	0.68	0.68	0.68	869
1	0.31	0.26	0.28	467
2	0.52	0.59	0.55	560
Accuracy		0.55		1896
Macro Avg	0.50	0.51	0.50	1896
Weighted Avg	0.54	0.55	0.54	1896

Table 3: Logistic Regression

Class	Precision	Recall	F1-score	Support
0	0.72	0.65	0.68	869
1	0.33	0.35	0.34	467
2	0.57	0.62	0.59	560
Accuracy		0.57		1896
Macro Avg	0.54	0.54	0.54	1896
Weighted Avg	0.58	0.57	0.57	1896

7.2 Confusion matrix

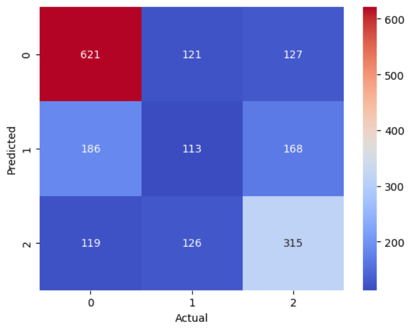

Figure 1: Random Forest

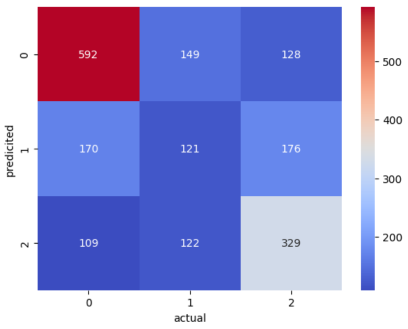

Figure 2: Neural Network

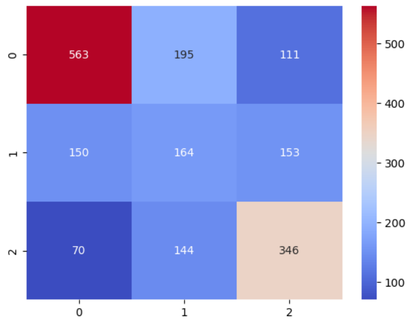

Figure 3: Logistic regression

The **Random Forest model** maximized the total number of correct predictions, while **Neural Network** remained slightly consistent in prediction for all classes.

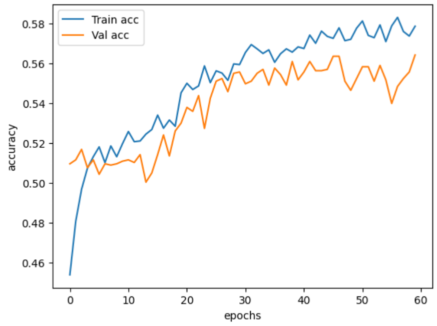

Figure 4: Accuracy progression over epochs

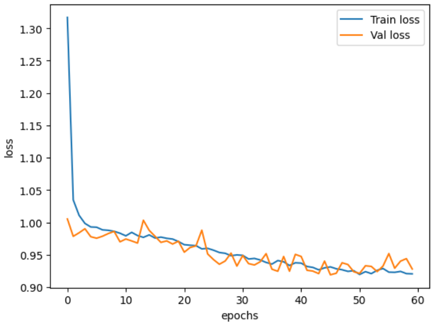

Figure 5: Loss minimization over epochs

7.3 Model Limitations

- **Random Forest:** Overpredicts draws; draw probability dominates many matches.
- **Neural Network:** Rarely predicts draws; mostly predicts home or away wins.
- **Logistic Regression:** Mostly predicts home wins and draws; underestimates away wins; sensitive to feature scaling and class imbalance.