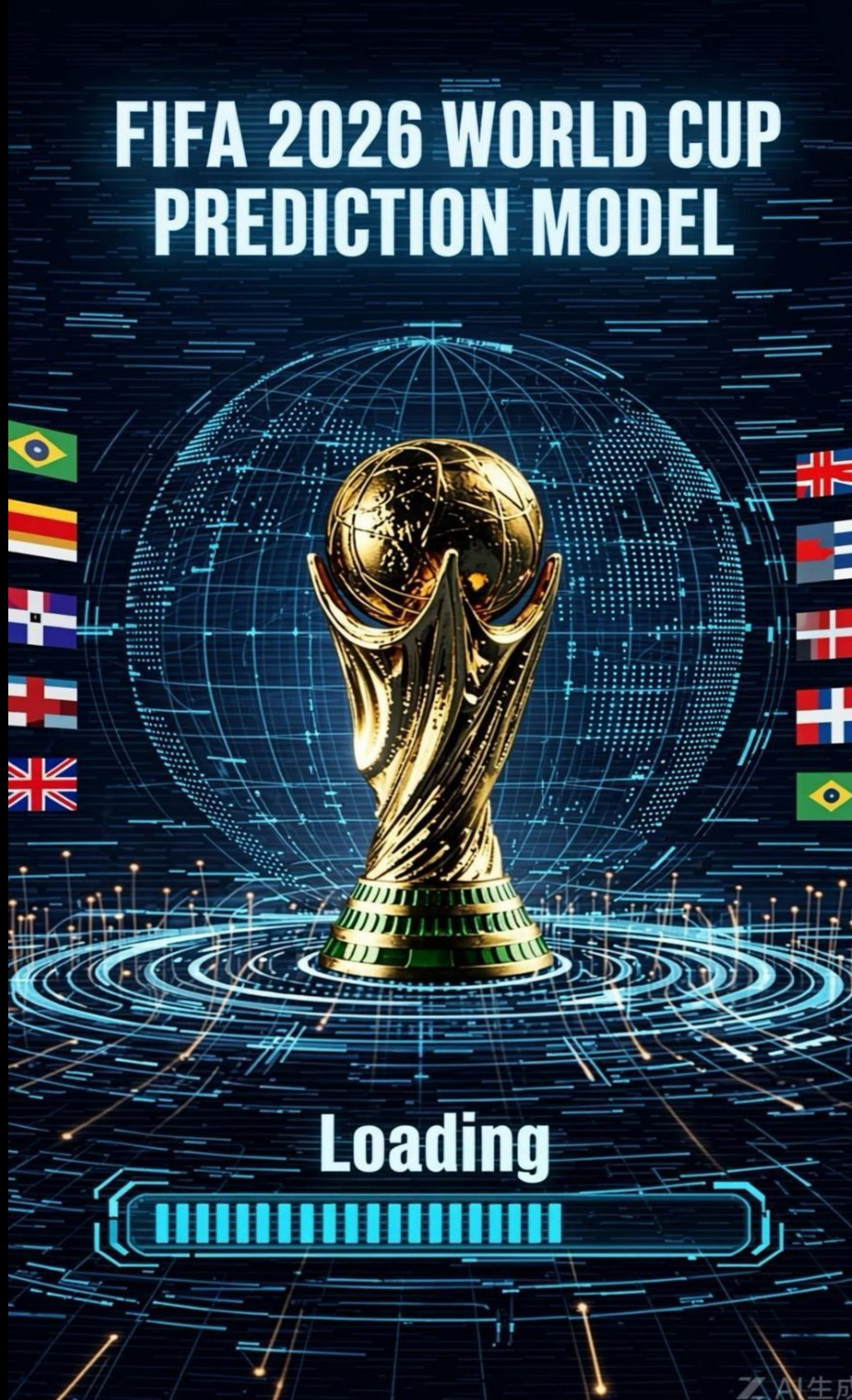

FIFA 2026 WORLD CUP
PREDICTION MODEL
Loading
AI生成

Predicting FIFA World Cup 2026 Outcomes

Probabilistic Modelling of Goals and Tournament Outcomes Using Statistical and Machine Learning Methods

DATA 606 Group Members:

- Tanvir Reza Chowdhury
- Taiseer Rakiin Ahad
- Chris Petrauskas

Core Questions

Who will win the tournament?

Who will win individual matches?

How many goals will be scored?

How far will teams advance?

Datasets

International Match Results

- Source: Public, community maintained dataset (Kaggle)
- Historical men's international matches
- Match date, teams, full-time score, venue
- 49,000 observations before processing

Official FIFA ranking releases

- Source: Fifa website
- Historical team strength ratings
- 68,000 observations before processing

Cleaning and Preparing Data

Preparation Steps

- Cleaned, removed and standardized missing values
- Removed friendly matches
- Merged nearest prior FIFA rankings to avoid look-ahead bias
- Filled missing ranking values using average
- Selected top 48 teams according to FIFA ranking points for 2026 world cup simulation

Key Features

- Fifa ranking difference
- Home advantage (only when not neutral)
- Neutral indicator
- Attack (team)/defence (opponent) relative strengths
- Recency weighting

$$\text{attack\_advantage} = \log\left(\frac{\text{team avg goals / global avg}}{\text{opponent avg conceded / global avg}}\right)$$

Two Modelling Approaches

1. **Goal-Based Model**
   Models expected goals using a GLM (Poisson/Negative Binomial)
   → Determine W/D/L from simulated goal counts

2. **Outcome-Based Model**
   Directly estimates W/D/L probabilities using ML classifiers
   → Sample match outcomes

Both approaches:
Run full tournament simulation with all 48 teams

Tournament Simulation

2026 Format Simulated

- 48 teams
- 12 groups of 4
- Top 2 groups + best 6 $\longrightarrow$ Round of 32
- Knockout to champion

Two simulation pipelines

- Goal-based (Negative Binomial sampling)
- ML-based (sample W/D/L from probabilities)
- We ran 2000 simulations and took averages

Why This Is Hard

- **Low-Scoring, High Variance Sport**
  - Goals are rare count events
  - Variance often exceeds the mean
  - One goal can flip outcome
  - Poisson assumptions may fail

- **Outcomes Are Inherently Probabilistic**
  - Upsets are common (a lot of games close to 50:50)
  - Draws are frequent
  - Home / neutral effects matter
  - Deterministic prediction is unrealistic

- **Tournament Structure Amplifies Noise**
  - 48 teams
  - Group + knockout stages
  - Small probability differences cascade
  - We must simulate distributions, not single outcomes

First Statistical Model

Poisson GLM (Count Model)

$$Y_i \approx Poisson(\mu_i)$$

$$\log(\mu_i) = \beta_0 + \beta_1 \cdot \text{points\_diff}_i + \beta_2 \text{ home\_adv}_i + \beta_3 \text{ neutral}_i + \beta_4 \text{ attack\_advantage}_i$$

$$\mu_i = \exp\left(\begin{pmatrix} \beta_0 + \beta_1 \text{points\_diff}_i + \beta_2 \text{ home\_adv}_i \\ + \beta_3 \text{ neutral}_i + \beta_4 \text{ attack\_advantage}_i \end{pmatrix} \right.$$

where: Y_i : goals scored
μ_i : expected goals

Fitted via maximum likelihood

Poisson assumption!
Variance ≈ Mean

Goal-based Model Diagnostic

Dispersion Ratio = Residual Deviance / Residual Degrees of Freedom
Overdispersion ratio ≈ **1.88** > 1

Interpretation

- Variance > Mean
- Poisson assumption violated

Solution

- Upgraded to **Negative Binomial GLM**
- Alpha ≈ 0.669
- Deviance dropped from 7,888 to 4556 - observed goal counts explained better
- (Alpha quantifies the degree of overdispersion, allowing the variance to grow faster than the mean, which better reflects the variability observed in goal counts)

$$\mathrm{Var}(Y_i) = \mu_i + \alpha\mu_i^2$$

This model was used for tournament goal simulation.

Model Results

- The explanatory variables ‘***points_diff***’, ‘***home_adv***’, ‘***attack_advantage***’ and ‘***neutral***’ were all strongly significant
- Higher ranking teams score more
- Home advantage and neutral venues increase goals scored
- Interestingly ‘***attack_advantage***’ DECREASES goals, possible explanations:
  - Collinearity with points_diff? Investigated and found unweighted correlation of 0.016
  - Attacking teams cancelling out?

Coefficient Estimates

Variable	Coef	Std Err	z	p-value	95% CI
Intercept	0.0119	0.031	0.383	0.701	[-0.049, 0.073]
points_diff	**0.0008**	4.84e-05	16.15	<0.001	[0.001, 0.001]
home_adv	**0.4608**	0.041	11.20	<0.001	[0.380, 0.541]
attack_advantage	**-0.1575**	0.028	-5.70	<0.001	[-0.212, -0.103]
neutral	**0.3057**	0.042	7.32	<0.001	[0.224, 0.388]

Machine Learning Statistical Model

Outcome Classification Model
Target: Win / Draw / Loss

Models

- Random Forest
- HistGradientBoosting
- Ensemble (average probabilities)

Training

- Stratified 5-fold CV
- RandomizedSearchCV
- Optimized for log-loss

Model metrics

Overall Performance

- Accuracy
- Macro F1 (treats all classes equally)

Probability Quality

- Log-loss (probability quality)
- Multiclass Brier score

Diagnostic Checks

- Confusion matrices
- Calibration plots

Key Model Findings

Probability Quality > Raw Accuracy

- Models optimized for log-loss, not just accuracy
- Ensemble delivered best log-loss (~1.025) and best Brier (~0.614)
- Calibration improved vs single models

Modest Predictive Signal

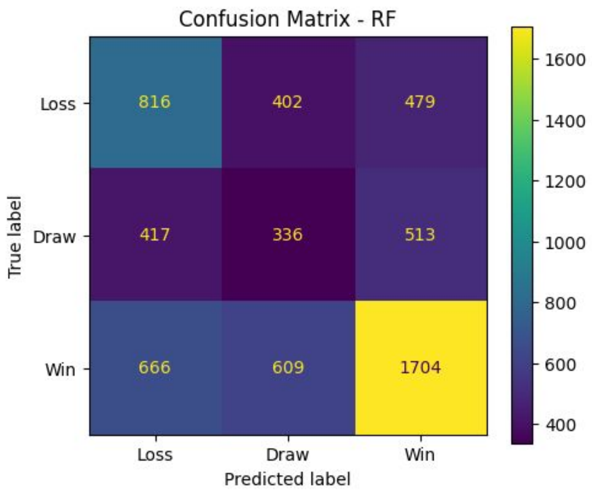
- Many matches close to 50:50 → inherent uncertainty
- Accuracy only moderately above baseline

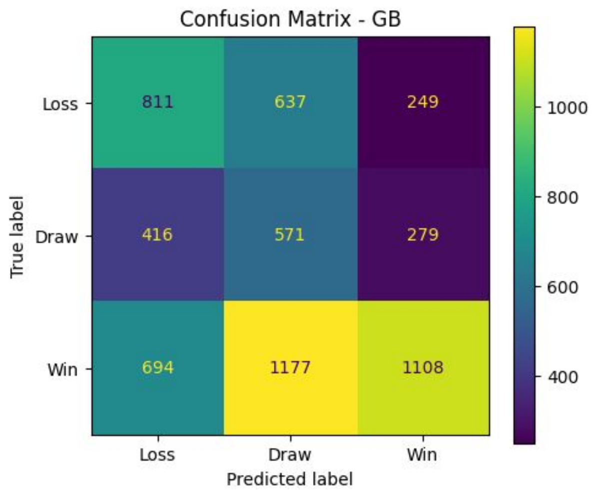
Feature Importance

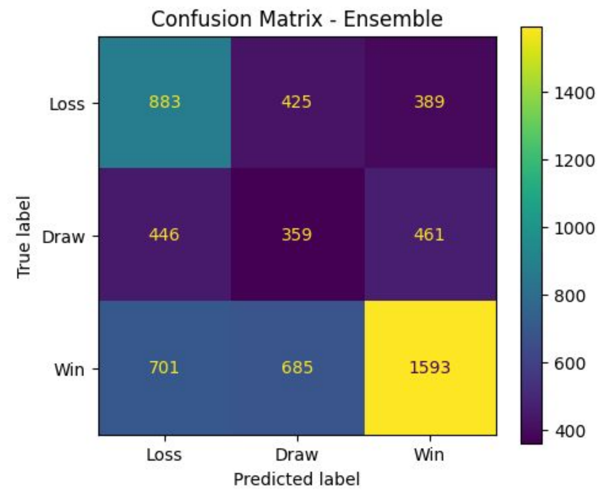
- FIFA points difference strongest predictor
- Home advantage meaningful
- Attack advantage smaller marginal impact
- Attack advantage had weak and non-monotonic effects in the ML models, consistent with it being a secondary signal rather than a dominant driver.

Draws Remain Hard to Predict

- Confusion matrices show draws most misclassified - reflects real-world parity and noise

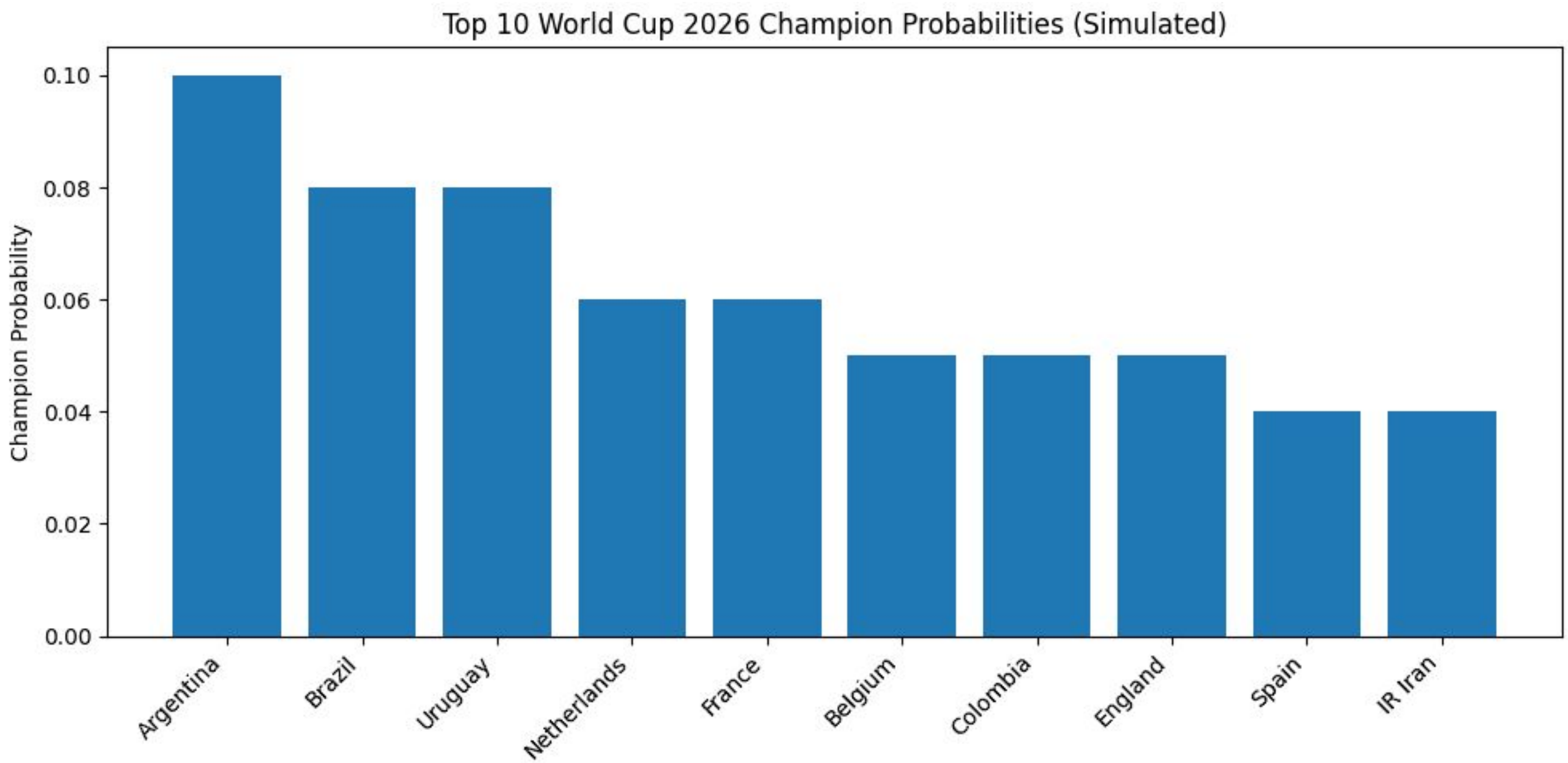






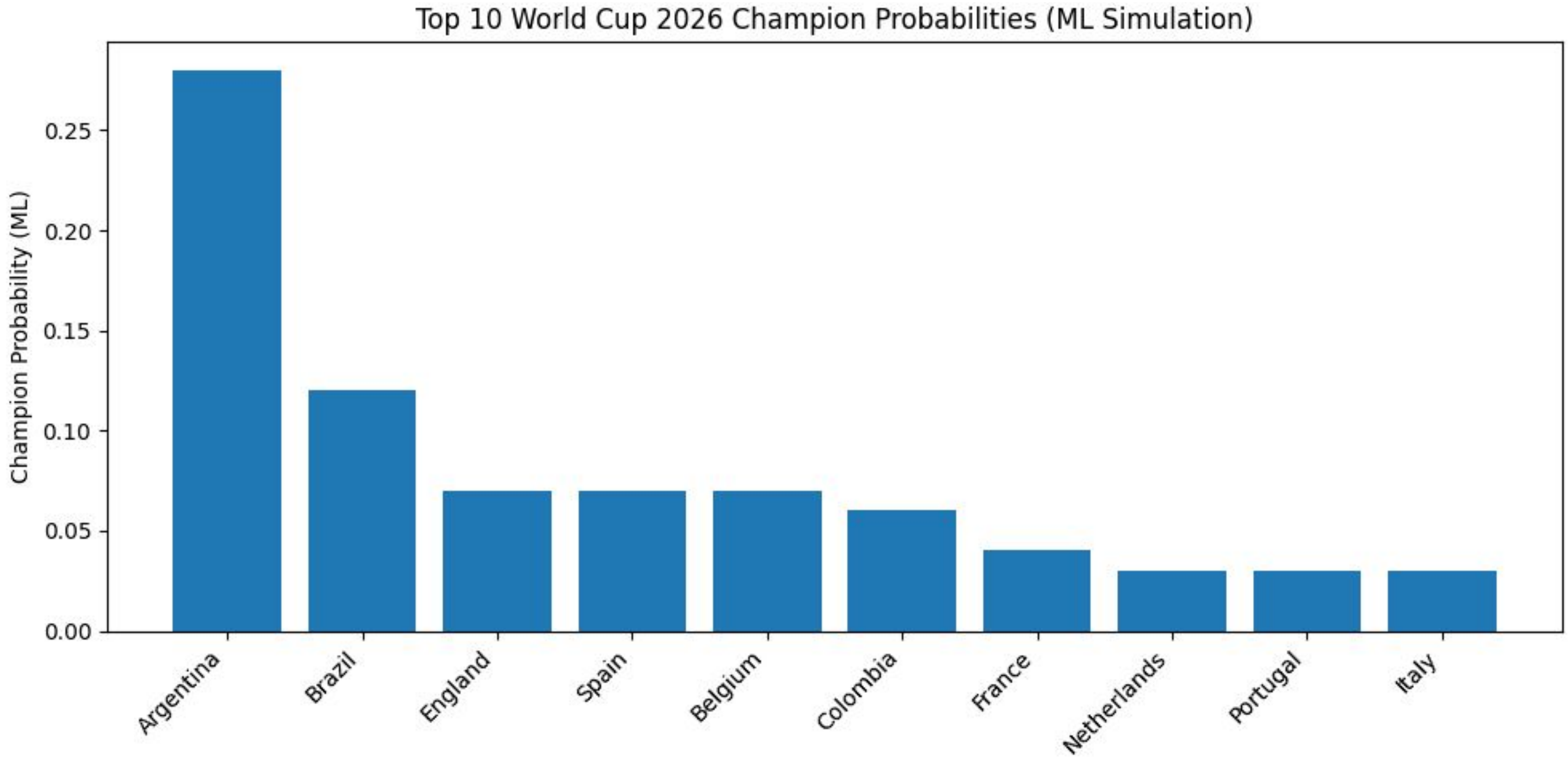
Simulation Results

Goal Model (GLM) Champions



Simulation Results

ML Model Champions (sample runs)



Assumptions & Limitations

- Results were plausible - but unstable
- FIFA rankings imperfect proxy for strength
- No injuries, tactics, or player-level data
- Simplified scorelines in ML simulation
- Potential temporal leakage if not time-split

Future Improvements

High-impact upgrades:

- Increase simulations (5,000–50,000)
- Time-based train/test splits
- Better score generation in ML simulation
- Take into account player rankings and penalties
- Team fixed effects ratings
- Calibrate the ensemble probabilities better

Thank you!
Any Questions?

2026

FIFA WORLD CUP

UNITED STATES - CANADA - MEXICO