# Predicting FIFA World Cup 2026 Outcomes

*Probabilistic Modelling of Goals and Tournament Outcomes Using Statistical and Machine Learning Methods (DATA 606)*

**Group Members:**

Taiseer Rakiin Ahad (30282008)

Tanvir Reza Chowdhury (UCID)

Chris Petrauskas (30270696)

## Abstract

This project builds and compares two different pipelines for predicting match outcomes via simulation of the FIFA World Cup 2026 format with the top 48 teams according to FIFA point rankings. The first model we used is a Generalized Linear Model (GLM). We start with a Poisson model and then upgrade to a Negative Binomial model after diagnosing overdispersion. For the second model, we train machine learning classifiers to predict Win/Draw/Loss probabilities and use those probabilities to simulate tournaments. We realized that accuracy alone may not be a reliable metric for determining how well the models are performing and so we decided to focus on probability-quality metrics such as log-loss and multiclass Brier score. Overall, FIFA ranking strength differences and home advantage were consistent predictors, while tournament outcomes remained noisy due to limited simulation runs.

## 1. Problem Definition and Motivation

Football is typically a low-scoring sport where randomness plays a big role. A single goal can change the entire outcome last minute and even strong teams can be eliminated in a knockout match. Because of this, predicting a tournament should not be treated as a single deterministic forecast. Instead, it is more realistic to build probabilistic models and simulate many possible tournament paths.

The core questions we focused on were:

- Who is most likely to win the 2026 World Cup?
- What are the probabilities of Win/Draw/Loss for individual matches?
- How many goals are expected to be scored in matches?
- How far are teams likely to advance in the tournament?

Our motivation for choosing this topic was because of 2 reasons. One is that the FIFA World Cup 2026 is fast approaching, and also the fact that probabilistic forecasting is widely used in real-world sports analytics as it supports decision-making under uncertainty. From a course perspective, this project is a good match for DATA 606 because it combines statistical modelling of numerical outcomes with classification-based prediction while also focusing on correct understanding of model assumptions and diagnostics.

*This project aims to develop a probabilistic framework for forecasting the FIFA World Cup 2026 by integrating statistical modeling, machine learning, and simulation. The specific objectives are:*

1. *Model goals scored using Poisson and Negative Binomial GLMs.*

2. *Predict match outcomes using machine learning classifiers.*

3. *Simulate the full 48-team tournament using Monte Carlo methods.*

4. *Estimate championship probabilities and quantify uncertainty.*

*This work aligns with the core learning objectives of DATA 606, including generalized linear modeling, model diagnostics, probability calibration, and simulation-based inference.*

## 2. Data Sources and Preparation

### 2.1 Datasets

| Dataset | What it contains / why we used it |
|---|---|
| **International match results** | 49071 rows, 9 columns (before processing). Contains match date, teams, score, tournament type, and neutral-site indicator. |
| **FIFA rankings** | 67894 rows, 6 columns (before processing). Contains historical FIFA points by team and ranking date. |

We used the match results to learn relationships between team strength indicators and match outcomes, and we used FIFA ranking points as a consistent proxy for team strength over time.

### 2.2 Cleaning and Feature Engineering

In order to make the datasets ready of training and analysis, we performed the following actions:

a) Standardized column names, parsed dates, and removed rows with missing critical values (teams, date, or score).

b) Removed friendly matches because they often have lower competitive intensity and can add noise when modelling tournament-level performance.

c) Converted the match dataset into a long format (one row per team, per match) to support goal modelling at the team level.

d) Time-aligned FIFA rankings to matches using the nearest prior ranking release (a backward merge), which helps avoid look-ahead bias issues.

e) Filled missing ranking values using the median ranking points as a reasonable fallback.

f) Computed a recency weight to give more value to recent matches:

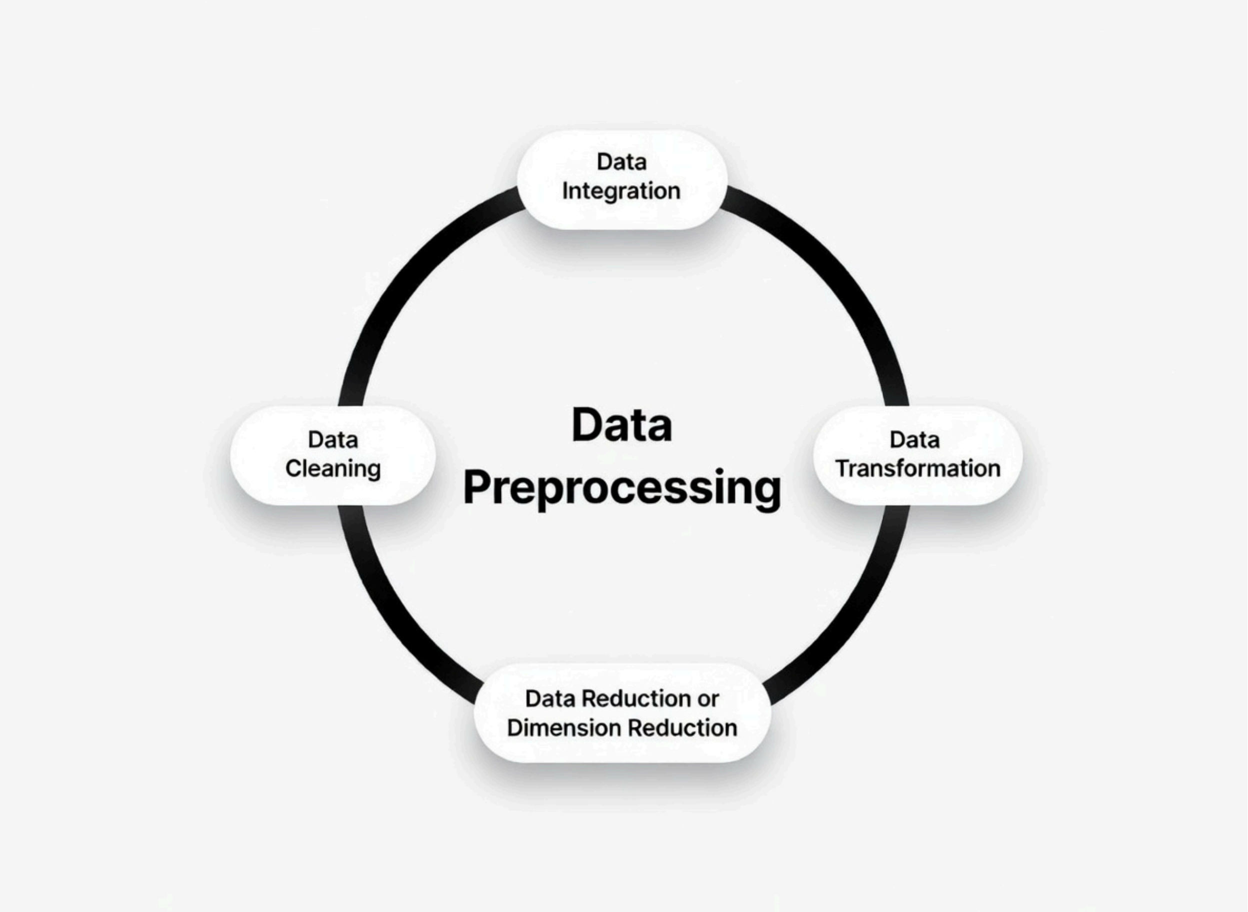
$$weight = 1 / (1 + years\_since\_match)$$

### Key engineered features:

| Feature | Definition |
|---|---|
| **points_diff** | Team FIFA points minus opponent FIFA points (a strength difference proxy). |
| **home_adv** | Home advantage indicator, applied only when the match is not neutral-site. |
| **neutral** | Neutral-site indicator (1 if neutral, 0 otherwise). |
| **attack_advantage** | log(attack_strength(team) / defence_strength(opponent)), where strengths are computed from average goals for/against relative to the global average. |
| **weight** | Recency weighting factor applied as sample weights in modelling. |

### 2.2.1 Modeling Considerations

*Special care was taken to avoid information leakage during data preparation. FIFA rankings were merged using the most recent ranking released prior to the match date to ensure that only information available at the time of the match was used. In addition, recency weighting was applied so that more recent matches had greater influence during model estimation. These steps help ensure that the models reflect realistic forecasting conditions rather than retrospective fitting.*



***Data processing flow diagram (raw data → cleaned data → long format → features)***

## 3. Methodology

We used two modelling strategies that answer similar questions in different ways:

1) The goal-based model predicts expected goals and then simulates scores.
2) The outcome-based model predicts Win/Draw/Loss probabilities directly.

### 3.1 Goal-Based Statistical Model

Since goals are non-negative, we used GLM as a starting point for this problem. We first fit a Poisson regression model:

*goals ~= points_diff + home_adv + attack_advantage + neutral*

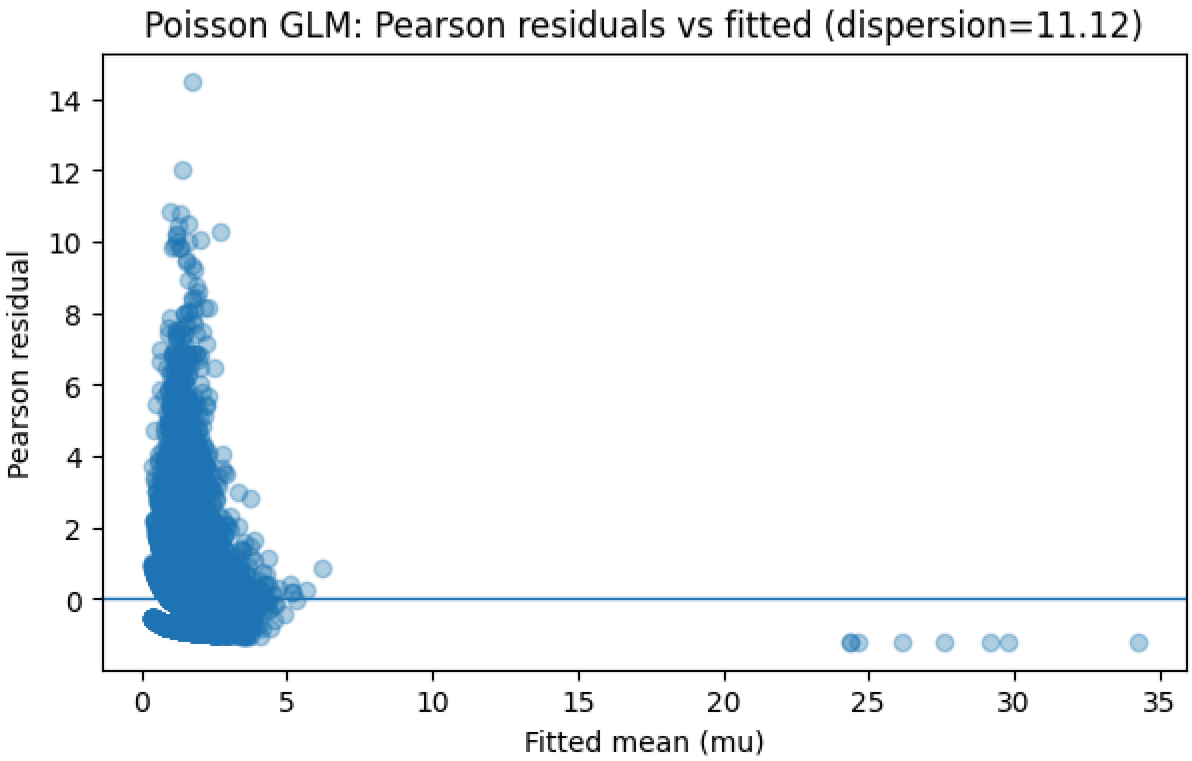
After fitting the Poisson model, we checked the Poisson variance assumption (variance ≈ mean) using an overdispersion diagnostic:

*Overdispersion ratio = residual deviance / residual degrees of freedom*

= 1.88 (> 1), indicating overdispersion.

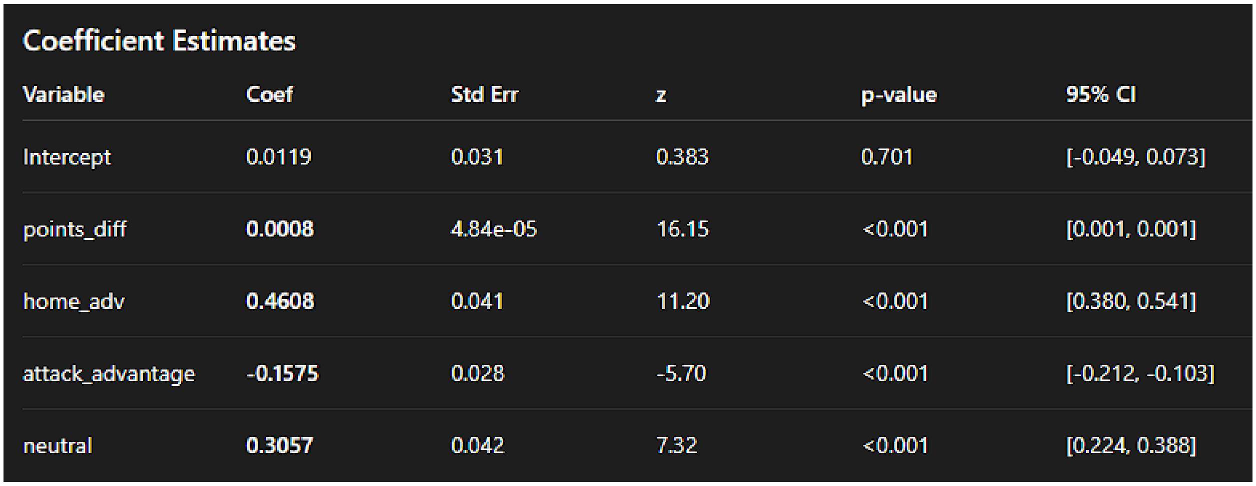
*The dispersion ratio greater than one indicates that the variance of goals exceeds the mean, violating the Poisson assumption of equidispersion. Under such conditions, the Poisson model underestimates variability and produces overly confident predictions. The Negative Binomial model introduces an additional dispersion parameter that allows the variance to exceed the mean, providing a more realistic representation of match uncertainty.*

Since overdispersion means the Poisson model underestimates variability, we upgraded to a Negative Binomial GLM. We estimated the dispersion parameter alpha ≈ 0.669310 and refit the model.



***Goal distribution / overdispersion diagnostic plot (histogram or mean vs variance)***

This actually improved the model fit and deviance dropped from about 7888.5 (Poisson) to 4556.6 (Negative Binomial).



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] |

***GLM coefficient summary table (Poisson vs Negative Binomial)***

## 3.2 Outcome-Based Machine Learning Model

For the second pipeline, we predicted match outcomes directly using classifiers. We defined the target from the home team's perspective: Loss (0), Draw (1), Win (2).

We used the same feature set as the GLM for consistency, and applied recency weights as sample weights. The models used were:

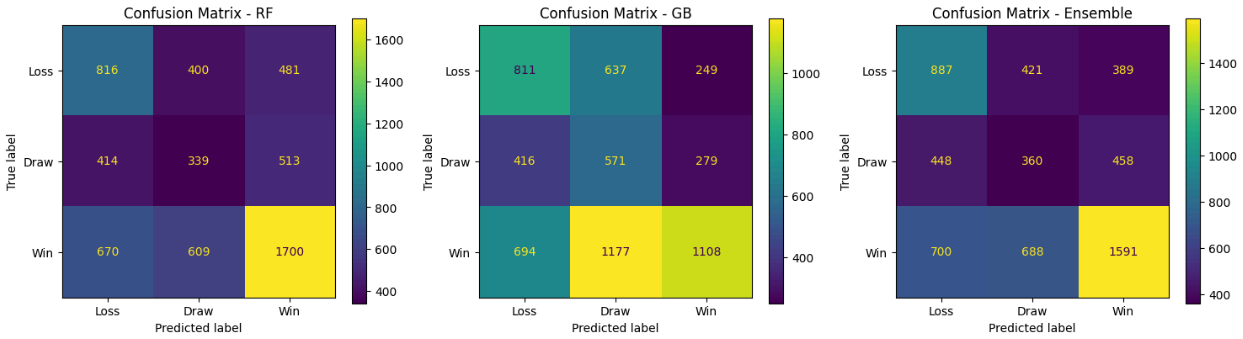
1) **Random Forest Classifier** because it handles non-linear interactions well.

2) **Histogram-based Gradient Boosting Classifier** since it has a strong performance on tabular data and requires limited feature engineering.

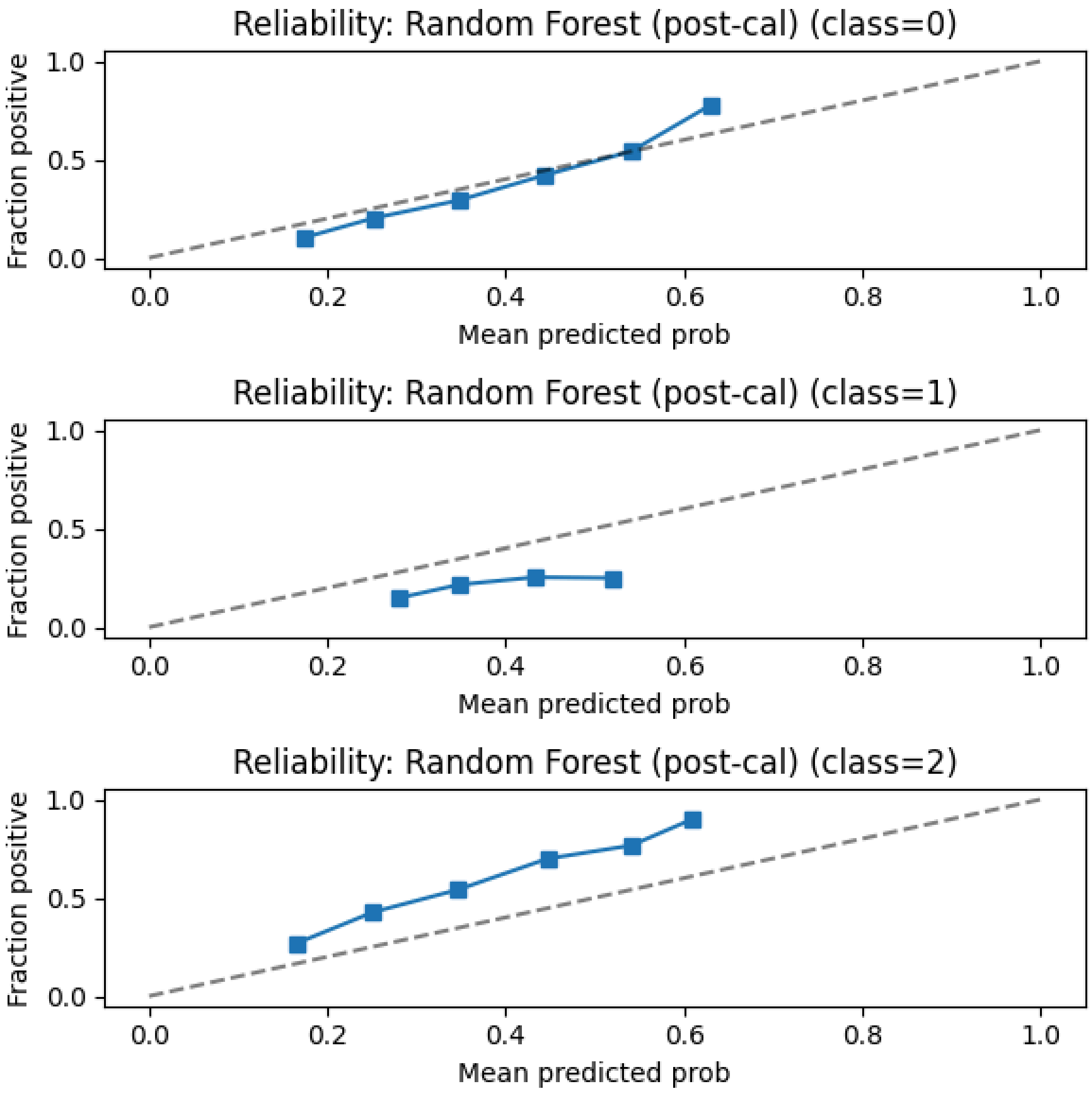
3) **Ensemble model**, which is an average of the predicted probabilities from Random Forest and Gradient Boosting.

For training and tuning the model we used stratified 5-fold cross-validation and RandomizedSearchCV, and optimized the model based on negative log-loss.

*Log-loss was selected as the primary optimization metric rather than accuracy because the tournament simulation relies on predicted probabilities rather than class labels. Well-calibrated probability estimates are therefore more important than raw classification performance.*



***Confusion matrix for best ML model (test set)***



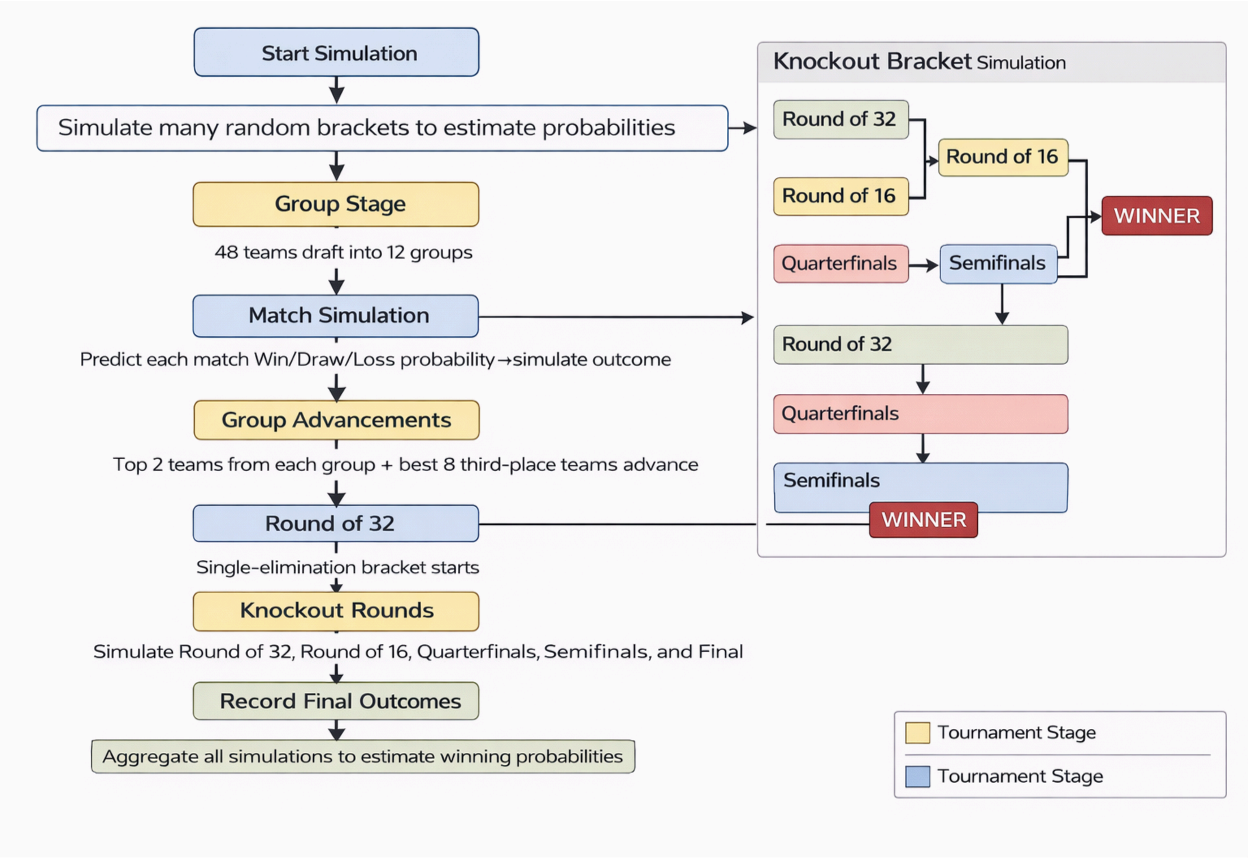
***Calibration (reliability) plot comparing models***

### 3.3 Tournament Simulation

We simulated the official 2026 format: 48 teams split into 12 groups of 4. From each group, the top 2 teams plus the best 6 third-place teams advance to a Round of 32, followed by standard knockout rounds until a champion is determined. In our submitted run, we used 2000 tournament simulations for each pipeline to estimate champion probabilities.

- **Goal-based pipeline:** predict expected goals for each team using the Negative Binomial GLM and sample goals to obtain match scores.
- **ML-based pipeline:** predict Win/Draw/Loss probabilities and sample an outcome. For group standings we used simplified scorelines, such as ‘1-0’, ‘0-1’, ‘1-1’.

Using probabilistic simulation rather than a single deterministic bracket allows the analysis to capture the inherent uncertainty of tournament outcomes. Small differences in match probabilities can propagate through the knockout structure, making Monte Carlo simulation essential for realistic tournament forecasting.


Start Simulation
Simulate many random brackets to estimate probabilities
Group Stage
48 teams draft into 12 groups
Match Simulation
Predict each match Win/Draw/Loss probability→simulate outcome
Group Advancements
Top 2 teams from each group + best 8 third-place teams advance
Round of 32
Single-elimination bracket starts
Knockout Rounds
Simulate Round of 32, Round of 16, Quarterfinals, Semifinals, and Final
Record Final Outcomes
Aggregate all simulations to estimate winning probabilities
Knockout Bracket Simulation
Round of 32
Round of 16
Round of 16
WINNER
Quarterfinals
Semifinals
Round of 32
Quarterfinals
Semifinals
WINNER
🟨 Tournament Stage
🟦 Tournament Stage

*Tournament bracket / simulation logic diagram*

# 4. Results

## 4.1 Goal Model Results

The goal model confirmed that FIFA strength differences and location effects do matter. In the Negative Binomial model, ‘*points_diff*’, ‘*home_adv*’, ‘*attack_advantage*’, and ‘*neutral*’ were statistically significant in our outputs.

| Predictor | Estimated direction (NB GLM) |
|---|---|
| **points_diff** | +0.0008 (log-link coefficient) |
| **home_adv** | +0.4608 (log-link coefficient) |
| **attack_advantage** | -0.1575 (log-link coefficient) |
| **neutral** | +0.3057 (log-link coefficient) |

- ***points_diff* > *0:*** stronger teams with higher FIFA points are expected to score more.

- ***home_adv* > *0:*** home advantage increases expected goals when the match is not neutral-site.

- ***neutral* > *0:*** neutral settings showed a positive coefficient in our historical sample.

- ***attack_advantage* < *0:*** this was a bit surprising, but the possible reasons include feature interactions, measurement noise in our simple attack/defence proxies or maybe some collinearity/overlap with points_diff.

## 4.2 ML Model Results

We evaluated classification models using both accuracy and probability-quality metrics. Accuracy alone is not enough because two models with similar accuracy can produce very different probability estimates.

### Baseline (simple train/test) accuracy

| Model | Accuracy |
|---|---|
| **RandomForest** | 0.491 |
| **HistGradientBoosting** | 0.542 |

### Tuned models (held-out test set metrics)

| Model | Test performance |
|---|---|
| **RandomForest** | Accuracy=0.480, Log-loss=1.044, Brier=0.623, Macro F1=0.437 |
| **HistGradientBoosting** | Accuracy=0.419, Log-loss=1.054, Brier=0.633, Macro F1=0.414 |
| **Ensemble** | Accuracy=0.478, Log-loss=1.025, Brier=0.614, Macro F1=0.442 |

The ensemble performed best overall in terms of probability quality (lowest log-loss and Brier score). This supports the idea that averaging probabilities can actually reduce variance and improve calibration compared to relying on a single model.

### Calibration attempt

We also tested calibration using sigmoid for the individual models. In our outputs, post-calibration did not seem to improve overall performance.

| Model | Calibration diagnostics |
|---|---|
| **Random Forest after calibration** | Log-loss=1.0620, Brier=0.6392 |
| **HistGB after calibration** | Log-loss=1.0535, Brier=0.6329 |

We also observed that draws were the hardest class to predict as they were the most misclassified. This does somewhat align with real football because draws often come from small random differences in low-scoring games.

## 4.3 Tournament Simulation Results

Using 2000 simulated tournaments, we estimated champion probabilities for each pipeline. But since the simulation count is relatively small, these probabilities should be treated as rough estimates.

Top champions (Goal-based NB GLM simulation)

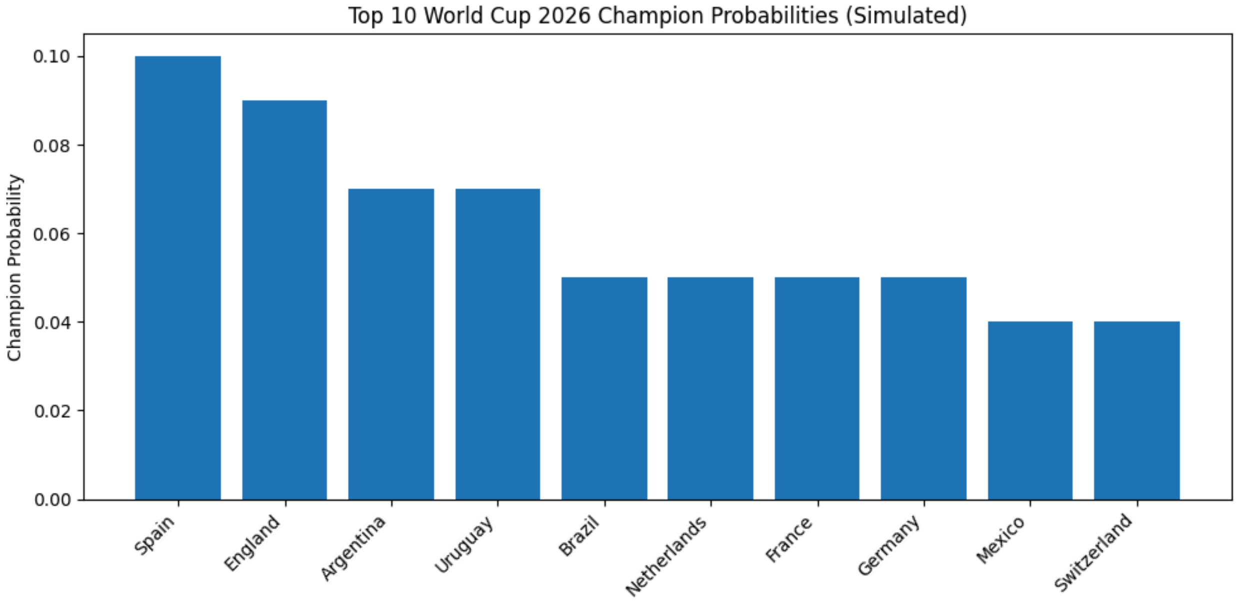
| Team | Champion probability |
|---|---|
| **Spain** | 0.10 |
| **England** | 0.09 |
| **Argentina** | 0.07 |
| **Uruguay** | 0.07 |
| **Brazil** | 0.05 |

Top champions (ML ensemble simulation)

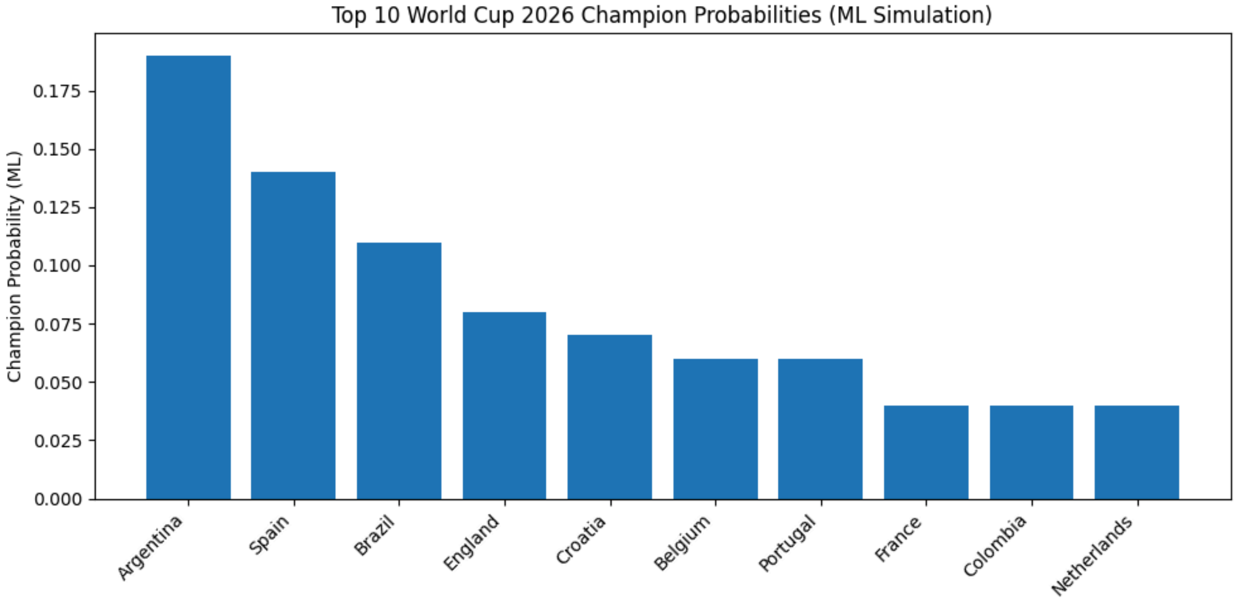
| Team | Champion probability |
|---|---|
| **Argentina** | 0.19 |
| **Spain** | 0.14 |
| **Brazil** | 0.11 |
| **England** | 0.08 |
| **Croatia** | 0.07 |

*An important observation is that both the goal-based and machine learning pipelines identified a similar group of leading contenders, including Argentina, Spain, Brazil, and England. The convergence of two methodologically different approaches increases confidence that the results reflect underlying team strength rather than model-specific bias.*

As an example of randomness in a single simulated tournament, one run produced Nigeria as champion. This is not a claim that Nigeria is likely to win, but it shows why we simulate distributions instead of trusting a single bracket.



***Bar chart of top 10 champion probabilities (Goal model)***



***Bar chart of top 10 champion probabilities (ML model)***

## 5. Discussion and Interpretation

Across both pipelines, FIFA ranking differences and location effects were consistently important. This matches what we initially expected: stronger teams tend to create and take advantage of more chances while home advantage can influence performance.

The Poisson-to-Negative-Binomial upgrade was a key modelling decision in our project. The overdispersion diagnosis showed that the Poisson model's core assumption was violated and that the Negative Binomial model better captured the extra variability in goals.

For match outcomes, the ML ensemble achieved the best probability quality. This matters because our tournament simulator relies on probabilities, and better-calibrated probabilities lead to more realistic simulated tournament distributions.

## 6. Assumptions and Limitations

Our main **assumptions** for this project were:

- FIFA ranking points are an adequate proxy for team strength.
- Matches were treated as conditionally independent given the features, since we did not take into account team strategy, injuries or penalties.
- Home advantage is simplified into a single indicator and only applied for non-neutral matches.
- Attack/defence strengths were computed from historical average goals as we did not have player level information available.

The **limitations** we faced for this project were as follows:

- Due to limited computational resources, the limited amount of tournament simulations made champion probabilities noisy. But increasing to much higher simulations would stabilize results.
- We used a random train/test split for ML evaluation. But a time-based split by training on earlier years and testing on later years would better reflect real forecasting.
- Attack/defense strengths were computed using aggregated match history rather than strictly training-only periods, which may have slightly inflated performance.
- The ML simulation used simplified scorelines, which reduced realism for goal difference and tie-break situations in group stages.
- No match context features, such as injuries, lineups, travel and match importance are included, so some important real-world factors are missing from the simulations.
- The machine learning evaluation used a random train–test split. A time-based split (training on earlier years and testing on more recent matches) would better reflect real forecasting conditions and reduce potential temporal leakage.

## 7. Conclusions

We successfully implemented and compared two methods taught in the DATA 606 course: a count-based GLM for goals using Poisson distribution and Negative Binomial, along with supervised ML classifiers for outcomes using Random Forest and boosting. Our diagnostics showed clear overdispersion in goals, so the Negative Binomial GLM was the more appropriate model for goal simulation. For outcome probabilities, the ensemble provided the best log-loss and Brier score, making it the strongest choice for probability-based tournament simulation.

Even though the estimates were sensitive to simulation randomness due to the low number of runs, overall we can say that the project demonstrates why probabilistic modelling with simulation is more honest and useful than a single predicted bracket.

*This project demonstrates how statistical modeling, machine learning, and simulation can be integrated to produce probabilistic forecasts for complex sporting events. The results highlight the importance of model diagnostics, probability calibration, and uncertainty quantification when making tournament-level predictions. Overall, the analysis illustrates the practical application of DATA 606 methods to a real-world decision-making problem.*

## 8. Future Improvements

For future improvements we would suggest:

- Increasing the tournament simulations to the range of 10,000-50,000 and report confidence intervals around champion probabilities.
- Use time-based splits and maybe rolling-window evaluation for more realistic forecasting.
- Improve ML simulation realism by generating scores conditional on outcomes.
- Adding more features such as match importance, travel distance, and player level data.

## References

- Kaggle: International football results dataset

```
'https://raw.githubusercontent.com/martj42/international_results/master/results.csv'
```

- FIFA: FIFA/Coca-Cola Men's World Ranking

```
'https://raw.githubusercontent.com/Dato-Futbol/fifa-ranking/master/ranking_fifa_historical.csv'
```

- Kaggle: FIFA World Ranking dataset