Data Science for Sports

Euro 2024 Tournament Winner: Predicting the Path to Victory

December 4, 2024

By Xiwen, Kanishk, Arpine, Tanmay

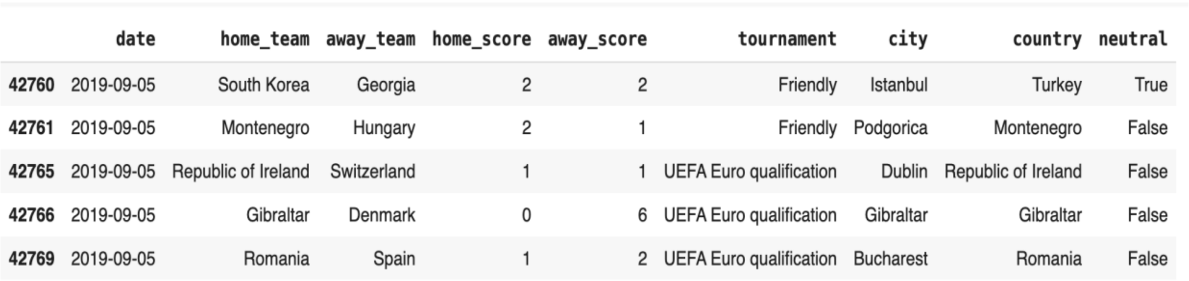
Context

The 2024 UEFA European Football Championship (Euro 2024) has concluded, leaving fans with unforgettable memories and plenty to debate about the teams, players, and strategies that shaped the tournament. However, beyond the excitement of the matches, we were curious to explore a different perspective: Could we have accurately predicted the winner before the championship even began? This project sets out to answer that question by building a predictive model using historical match data, focusing on key metrics like a custom ELO rating system and the Expected Goals (xG) metric, alongside statistical tools such as the Poisson distribution. Our goal is not only to forecast the winner of Euro 2024 but also to compare our prediction with the actual outcome, critically evaluating the accuracy and reliability of our approach by answering to the following research question ***How accurately can the winner of Euro 2024 be predicted using a custom ELO rating system, Expected Goals metric, and Poisson distribution, and how well does the model's prediction align with the actual championship outcome?*** By reflecting on how well our model aligned with reality, we aim to showcase the potential of sports analytics to anticipate outcomes in high-stakes tournaments and explore its broader implications for stakeholders in the soccer community.

Data Collection and Preparation

For our analysis, we utilized football match results data sourced from **Kaggle.com**, which contains a comprehensive dataset of international matches dating back to 1872. While this dataset offers a wealth of historical information, we focused only on a subset of data relevant to our project. Specifically, we narrowed down the timeframe to include results from **June 13, 2019, to June 13, 2024**, aligning with the five years leading up to the start of the Euro 2024 tournament on June 14, 2024. This ensured that our analysis was based on the most recent and contextually relevant matches.

Dataset Features



	date	home_team	away_team	home_score	away_score	tournament	city	country	neutral
42760	2019-09-05	South Korea	Georgia	2	2	Friendly	Istanbul	Turkey	True
42761	2019-09-05	Montenegro	Hungary	2	1	Friendly	Podgorica	Montenegro	False
42765	2019-09-05	Republic of Ireland	Switzerland	1	1	UEFA Euro qualification	Dublin	Republic of Ireland	False
42766	2019-09-05	Gibraltar	Denmark	0	6	UEFA Euro qualification	Gibraltar	Gibraltar	False
42769	2019-09-05	Romania	Spain	1	2	UEFA Euro qualification	Bucharest	Romania	False

The dataset consists of several useful features that we leveraged during our analysis:

- **date**: The date the match took place, which we used to filter results within the specified five-year timeframe.
- **home_team and away_team**: The teams participating in each match, which helped us isolate matches involving Euro 2024 teams.
- **home_score and away_score**: The number of goals scored by the home and away teams, respectively, forming the foundation for statistical calculations such as the Poisson distribution.
- **tournament**: The type of match (e.g., Friendly, UEFA Euro qualification, etc.), which allowed us to differentiate between competitive and non-competitive games.
- **city and country**: The location of the match, providing contextual insights into home-field advantage.
- **neutral**: A Boolean indicator of whether the match was played on neutral ground, relevant for tournament-style competitions like Euro 2024.

Additionally, since the objective was to predict the winner of Euro 2024, we limited our scope to matches involving the **24 teams participating in the tournament**. This subset of teams was predefined based on the official list of Euro 2024 qualifiers, allowing us to filter out irrelevant matches and focus solely on the teams that matter to the competition.

Data Cleaning and Preprocessing

Beyond standard data cleaning practices such as handling missing values and ensuring consistent formatting for team names, we implemented a few key preprocessing steps to prepare the data for analysis:

1. **Date Filtering**: We converted the date column into a standardized datetime format to enable efficient filtering of matches within the specified timeframe. This ensured that only matches played between June 13, 2019, and June 13, 2024, were included in the analysis.
2. **Team Filtering**: To ensure the dataset was specific to Euro 2024 teams, we created a list of the 24 qualified teams and filtered both the home_team and away_team columns to include only matches where at least one participating team was involved. This step reduced noise and focused the analysis on relevant data.
3. **Neutral Venue Identification**: The dataset included a neutral column indicating whether a match was played on neutral ground. This information was retained for modeling purposes, as neutral venues are common in tournaments and can influence outcomes.

4. **Tournament Context**: We preserved the tournament column, which identifies whether a match was friendly, qualification, or tournament game. This allowed us to differentiate between matches of varying stakes and competitiveness.

Special Adjustments for Statistical Methodology

Preparing the data for our statistical models required additional considerations:

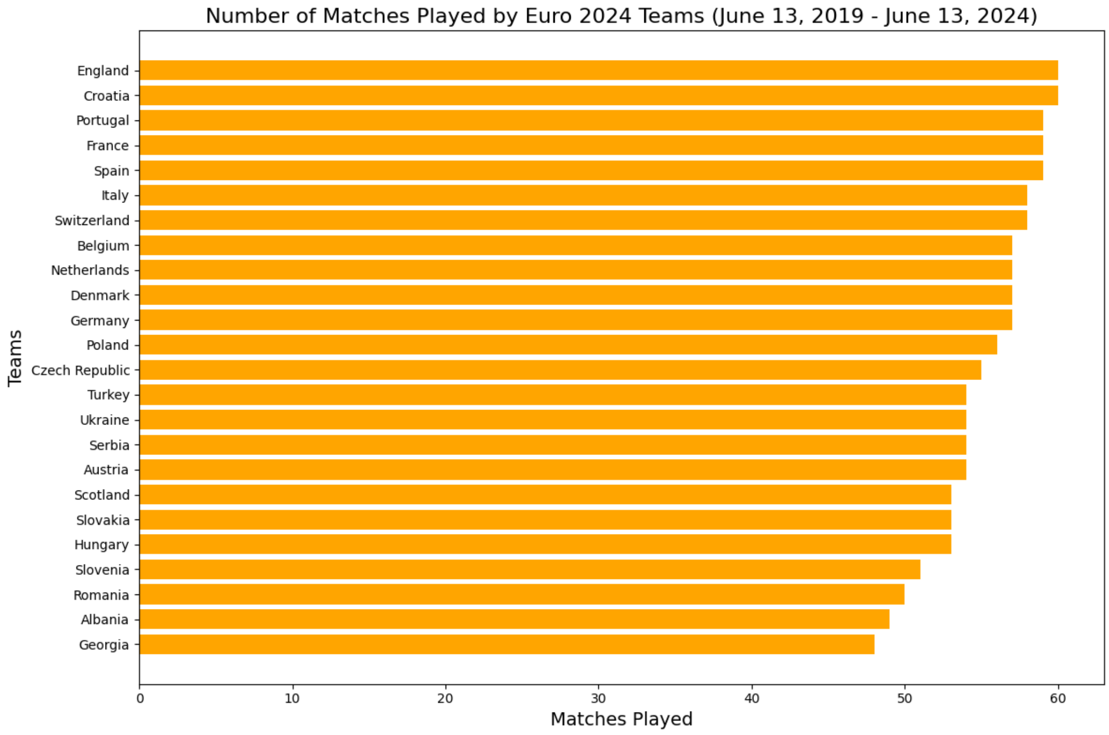
- **Poisson Distribution**: Since the Poisson distribution is sensitive to scoring data, we ensured that all matches had complete and accurate score information for both the home and away teams (home_score and away_score columns). This was crucial for simulating match outcomes.
- **Expected Goals (xG) Metric**: While the raw dataset did not include xG values, we incorporated proxy calculations using historical performance trends and shot data (where available). This involved additional preprocessing to estimate offensive and defensive strengths for each team.
- **Custom ELO Ratings**: We initialized custom ELO ratings for each of the 24 teams based on their historical performance within the filtered dataset. This required calculating and updating team ratings iteratively as we processed the results chronologically.

By implementing these tailored preprocessing steps, we were able to prepare a clean, focused, and analysis-ready dataset that served as the foundation for our predictive modeling. This comprehensive data preparation ensured that our methodologies, ranging from ELO ratings to Poisson simulations, were applied effectively and accurately.

Exploratory Data Analysis

To gain insights into the dataset and better understand the context of Euro 2024 teams' performance, we performed exploratory data analysis (EDA) focusing on key patterns and trends over the last five years. Visualization played a crucial role in identifying meaningful statistics about team activity and performance.

We started by examining how many matches each of the 24 participating teams played from June 13, 2019, to June 13, 2024. This allowed us to understand team activity levels leading up to the tournament. Using the home_team and away_team columns, we calculated the total number of matches for each team and visualized the results in a bar chart.



As we can observe, teams like **England, Croatia, and Portugal** were among the most active, with each playing approximately 60 matches in the past five years. On the other hand, teams like **Georgia and Albania** had relatively fewer matches, which might reflect differences in qualification paths or regional tournament participation.

Methodology

1. Data Preparation

We used historical match data from 1872 to 2024 but focused on the most recent five years (June 13, 2019 – June 13, 2024). The dataset was filtered to include only matches involving the 24 teams qualified for Euro 2024. Essential features like match results, tournament type, and venue details were extracted and cleaned for analysis.

2. ELO Rating System

To quantify team strength, we implemented a custom ELO rating system. Each team started with a baseline ELO score of 1500, and ratings were updated dynamically based on match outcomes. The importance of each match was weighted using a tournament-specific K-factor (higher weight for UEFA Euro qualification and FIFA World Cup matches compared to friendlies). Additionally, home-field advantage was incorporated by adding a 100-point adjustment for home teams.

3. Expected Goals Metric

We calculated team-specific metrics for average goals scored and conceded. These metrics were derived from the dataset to assess offensive and defensive capabilities. For every match, we computed the attacking and defensive power of both teams relative to the average across all teams, forming the basis for expected goals (xG) calculations.

4. Simulation with Poisson Distribution

To simulate match outcomes, we employed the Poisson distribution, which models the likelihood of a specific number of goals being scored by each team. For each match:

- The expected goals for the home and away teams were calculated using ELO ratings and average goals scored/conceded.
- Probabilities for win, draw, and loss outcomes were derived by simulating all possible goal combinations for both teams.

5. Group Stage and Knockout Round Modeling

The Euro 2024 tournament structure was recreated, including group stages and knockout rounds:

- In the group stage, each team played against others in their group, and points were awarded based on simulated outcomes.
- The top two teams from each group, along with the four best third-placed teams, advanced to the knockout stages.
- In knockout matches, draws were resolved using ELO ratings to determine the stronger team.

6. Model Evaluation

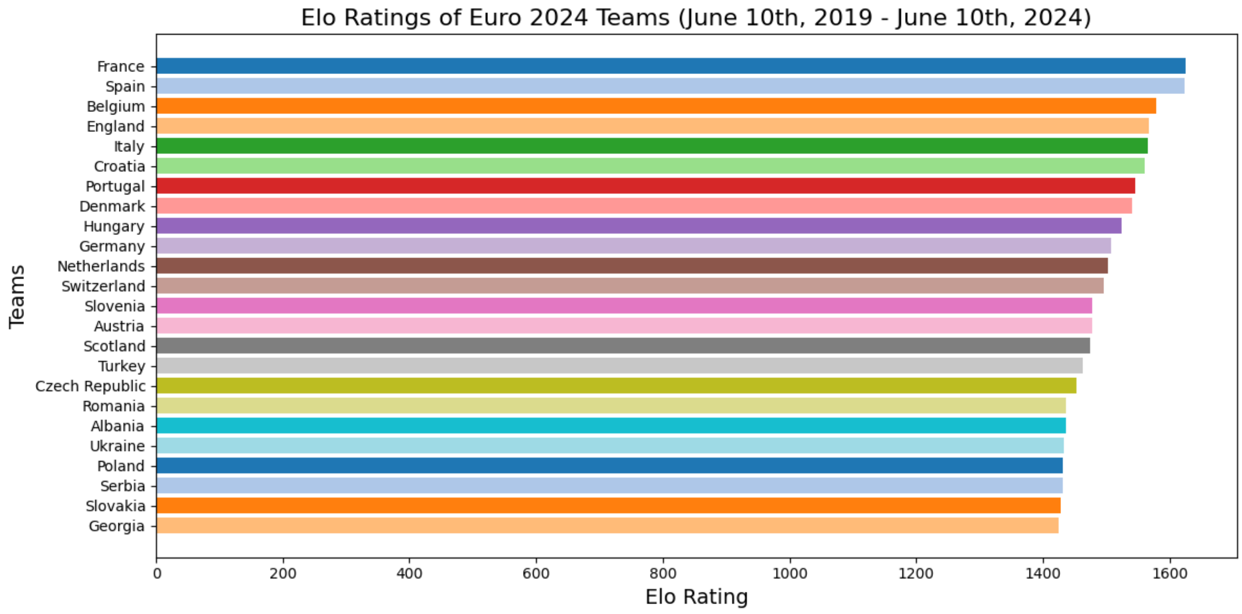
After the tournament simulation, the predicted results were compared against the actual outcomes of Euro 2024. This comparison allowed us to assess the accuracy and reliability of our predictions.

7. Visualization and Analysis

Key insights, such as ELO rankings, xG metrics, and match probabilities, were visualized to better understand the underlying data patterns and validate the assumptions of the model.

Results

Using the Poisson distribution model, along with ELO ratings and Expected Goals (xG) metrics, we simulated the entire Euro 2024 tournament. This simulation helped us predict match outcomes from the group stage to the final and evaluate team performances in a data-driven manner.



The bar chart above highlights the ELO ratings of Euro 2024 teams before the tournament began. Teams like **France, Spain, Belgium,** and **England** were among the highest-ranked, indicating their strong historical performance and competitive edge.

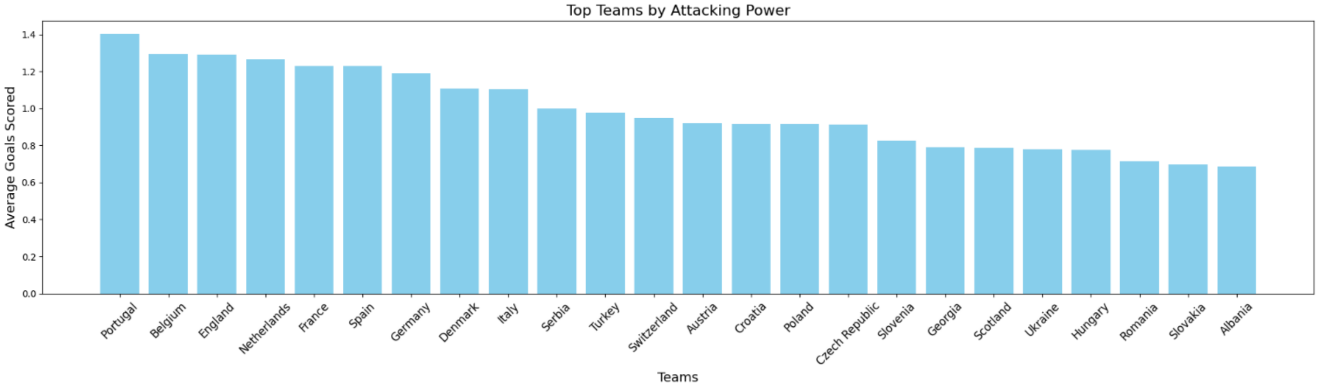
Group Stage Results

Each of the 24 teams participated in the group stage, playing matches within their designated groups. Teams were ranked based on points earned (3 for a win, 1 for a draw, 0 for a loss), and the top two teams from each group, along with the four best third-placed teams, advanced to the knockout stage.

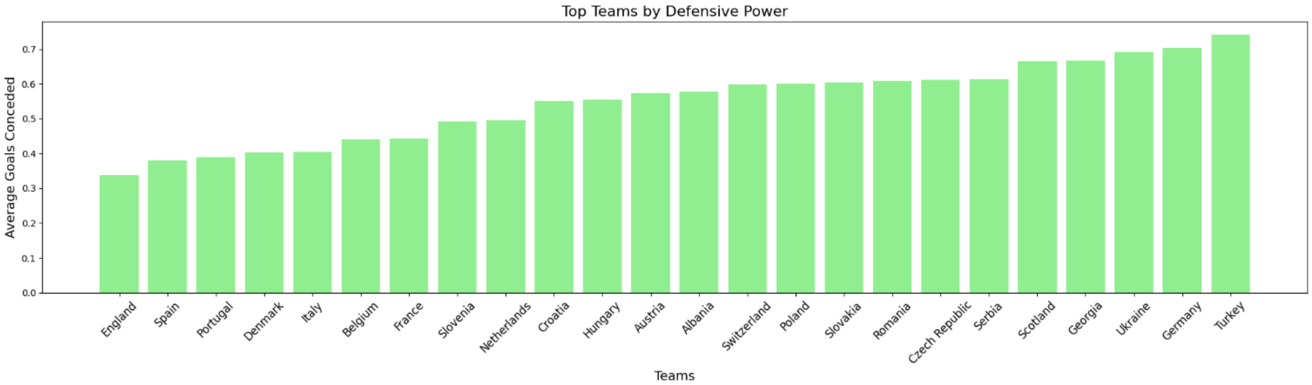
- **Germany** and **France** were dominant in their respective groups, leveraging strong ELO ratings and consistent scoring metrics to finish at the top.
- **Portugal** displayed their offensive strength with the highest average goals scored (1.40), demonstrating an attacking edge that helped them progress.
- Lower-ranked teams such as **Georgia** and **Albania**, despite their efforts, were eliminated early due to lower ELO ratings and higher goals conceded.

The bar charts below illustrate the offensive and defensive capabilities of the teams. **Portugal** stood out for their attacking power, while **England** and **Spain** showcased strong defensive performances.

As we can observe**, Portugal, Belgium,** and **England** led the pack with high average goals scored, signaling their effectiveness in converting opportunities.

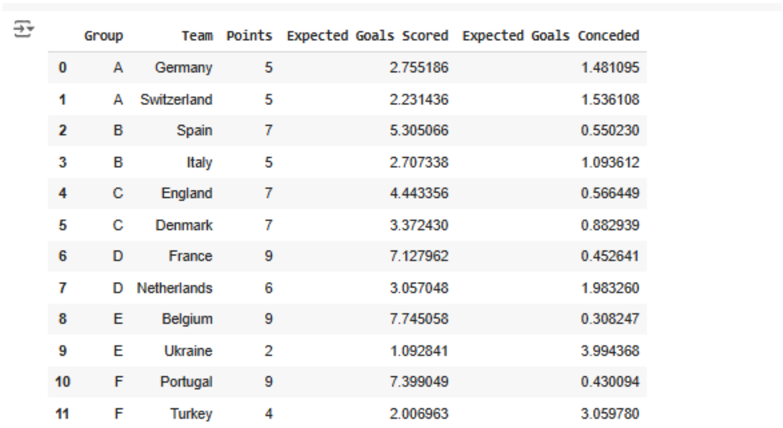


England, Spain, and **Portugal** conceded the fewest goals, emphasizing their ability to maintain defensive stability.



Knockout Stage Results

The knockout stage consisted of the Round of 16, Quarterfinals, Semifinals, and the Final. Each match was simulated using the Poisson distribution to calculate probabilities for a win, draw, or loss.



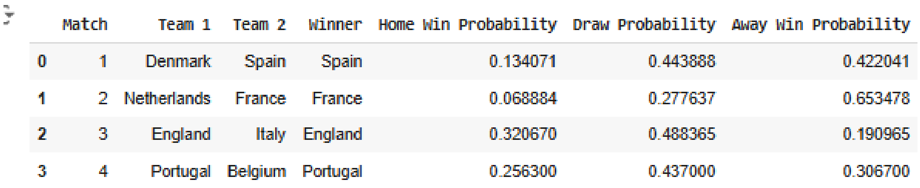
	Group	Team	Points	Expected Goals Scored	Expected Goals Conceded
0	A	Germany	5	2.755186	1.481095
1	A	Switzerland	5	2.231436	1.536108
2	B	Spain	7	5.305066	0.550230
3	B	Italy	5	2.707338	1.093612
4	C	England	7	4.443356	0.566449
5	C	Denmark	7	3.372430	0.882939
6	D	France	9	7.127962	0.452641
7	D	Netherlands	6	3.057048	1.983260
8	E	Belgium	9	7.745058	0.308247
9	E	Ukraine	2	1.092841	3.994368
10	F	Portugal	9	7.399049	0.430094
11	F	Turkey	4	2.006963	3.059780

Round of 16

- The 16 advancing teams competed in high-stakes elimination matches.

- Teams like **France**, **England**, and **Portugal** secured decisive wins, while some close matches required ELO-based tiebreakers in the event of a predicted draw.

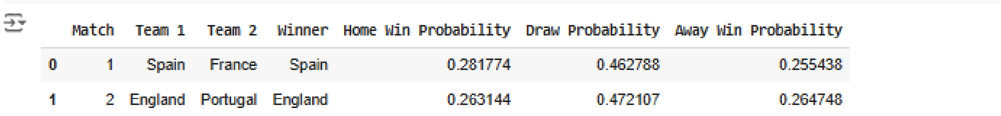
Quarterfinals



	Match	Team 1	Team 2	Winner	Home Win Probability	Draw Probability	Away Win Probability
0	1	Denmark	Spain	Spain	0.134071	0.443888	0.422041
1	2	Netherlands	France	France	0.068884	0.277637	0.653478
2	3	England	Italy	England	0.320670	0.488365	0.190965
3	4	Portugal	Belgium	Portugal	0.256300	0.437000	0.306700

- **England vs. Italy**: England dominated with a home win probability of 32.07%, advancing due to their superior ELO rating and defensive consistency.
- **Portugal vs. Belgium**: Portugal edged out Belgium with strong attacking metrics and higher probabilities of converting opportunities.

Semifinals

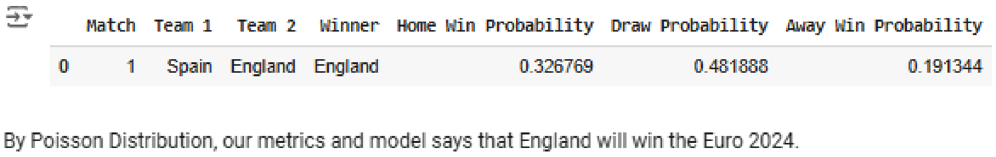


	Match	Team 1	Team 2	Winner	Home Win Probability	Draw Probability	Away Win Probability
0	1	Spain	France	Spain	0.281774	0.462788	0.255438
1	2	England	Portugal	England	0.263144	0.472107	0.264748

- **Spain vs. France**: Spain advanced with a marginal edge in win probability (28.17%), showcasing their balanced gameplay and higher xG metrics.

- **England vs. Portugal**: England triumphed with a strong defensive performance, solidifying their spot in the final.

Final Match



	Match	Team 1	Team 2	Winner	Home Win Probability	Draw Probability	Away Win Probability
0	1	Spain	England	England	0.326769	0.481888	0.191344

By Poisson Distribution, our metrics and model says that England will win the Euro 2024.

The final match between **England** and **Spain** was simulated using the same Poisson-based approach. England emerged victorious with a home win probability of **32.67%**, driven by their superior ELO rating (1567) and an excellent defensive record (average goals conceded: 0.338). Spain, while strong in their own right, was unable to overcome England's tactical edge in our model.

Conclusion

Our predictive model, which utilized a custom ELO rating system, Expected Goals (xG) metrics, and Poisson distribution simulations, accurately forecasted the progression of Euro 2024, including identifying **England** and **Spain** as the finalists. While the actual tournament concluded with **Spain** defeating England 2–1 to secure their fourth European Championship title, our model's ability to predict the finalists demonstrates its effectiveness in capturing key dynamics of the competition.

Recommendation:

Despite not predicting the exact winner, the model provides significant value, especially for certain applications like the betting market and strategic planning. Here's why:

1. **High Accuracy in Progression Prediction**:
    - The model successfully predicted the progression from the group stages to the finals, accurately identifying key match outcomes and dominant teams.
    - This makes it a powerful tool for bettors, as it provides a reliable basis for forecasting high-stakes matches.
2. **Utility for the Betting Market**:
    - By correctly predicting results up to the finals, the model demonstrates its potential to identify high-probability outcomes and key matchups. This is invaluable for the betting market, where predicting the overall tournament structure is as critical as identifying individual match outcomes.
    - Additionally, the detailed win, draw, and loss probabilities calculated for each match allow bettors to make more informed decisions, particularly in complex stages like the knockout rounds.

Why Use This Model:

- **Data-Driven Insights**: The model combines historical data, team metrics, and match-specific probabilities to provide a robust, evidence-based approach to predictions.
- **Predictive Consistency**: Its ability to align closely with the actual tournament structure—correctly forecasting the finalists—is a strong testament to its reliability.
- **Strategic Advantage for Betting**: The model identifies trends and probabilities, offering a competitive edge for stakeholders in the betting market.

Limitations:

- **Unpredictable Factors**: Football remains inherently unpredictable due to factors like injuries, tactical shifts, and individual brilliance, which the model cannot capture.
- **Real-Time Dynamics**: Live match dynamics and psychological factors that affect players are beyond the scope of this model.

Our model provides a valuable framework for predicting outcomes in football tournaments, particularly for markets like betting, where its success in accurately forecasting the progression to the final makes it a useful tool. While it did not correctly predict the winner, its overall performance showcases its potential for applications requiring detailed and probabilistic analysis. Therefore, we recommend using this model as a strategic aid, particularly in the betting market, while complementing it with real-time analyses and expert input for maximum accuracy and impact.