Predicting FIFA World Cup Match Outcomes Using Historical Data and Poisson Regression

Project Report

1 Objective

The primary goal of this project is to create a model to predict the outcomes of FIFA World Cup matches. The model leverages historical data from past tournaments to assess the strengths of teams and forecast match results using Poisson regression.

2 Data Collection

Data from all FIFA World Cups since 1930 was collected from Wikipedia. The project scrapes match results and team statistics from tournament pages. The key steps include:

- **Web Scraping:** Using `requests` and `BeautifulSoup`, matches from Wikipedia pages for each World Cup were scraped.
- **Data Storage:** The data is stored in pandas DataFrames, and intermediary tables are serialized using `pickle`.

3 Data Processing

The project focuses on cleaning and organizing the match data to make it usable for predictive modeling:

- **Cleaning Country Names:** Spaces and Unicode characters (e.g., en-dash) in team names were removed.
- **Score Cleaning:** Scores are represented as digits and hyphens by removing unnecessary characters, such as "After Extra Time" notations.
- **Goals Splitting:** Home and away goals are extracted into separate columns.
- **Renaming Columns:** Columns were renamed for better readability and compatibility with the model, such as renaming teams to `HomeTeam` and `AwayTeam`.

4 Team Strength Calculation

To predict outcomes, the project calculates the offensive and defensive strength of each team using historical data:

- **Splitting Data:** Match data is split into two DataFrames—one for home matches and one for away matches.
- **Goals Analysis:** Goals scored by each team and goals conceded by opponents were used to estimate the average performance of each team.
- **Team Strength Calculation:** Home and away data are combined to calculate the mean goals scored and conceded for each team, providing insights into overall team strength.

5 Poisson Regression Model

The core of the prediction model is based on the Poisson distribution:

- **Lambda Calculation:** For each match, the average number of goals each team is expected to score (λ_{home} and λ_{away}) is calculated using their historical strength.
- **Outcome Probabilities:** The probability of different match outcomes (win, draw, loss) is determined by calculating the likelihood of different goal combinations (0–10 goals for both home and away teams) using the Poisson probability mass function (PMF).
- **Points Prediction:** The project predicts the points for both home and away teams based on the calculated probabilities.

6 Fixtures Prediction for 2022 World Cup

The project predicts the outcomes for the group stage and knockout rounds of the 2022 FIFA World Cup:

- **Group Stage Simulation:** The match results for each group are simulated, and the points for each team are updated based on the Poisson model predictions.
- **Knockout Stages:** The winners of the group stage are used to fill the knockout rounds (Round of 16, Quarterfinals, Semifinals, and Finals). A function was created to simulate and update the winner of each knockout match.

7 Automation Functions

Two key functions automate the prediction process:

- `get_winner(df_fixture_updated)`: Predicts the winner of a match based on the teams' strengths and updates the fixture DataFrame.
- `update_table(prev_fixture, next_fixture)`: Transfers the winners from one knockout round to the next, ensuring accurate progression through the tournament.

8 Results

The project generates two primary outputs:

- **Historical Data:** The cleaned data from previous World Cups is saved in `FIFA_World_Cups_Historical_Data_Cleaned.csv`.
- **2022 World Cup Predictions:** The predicted fixtures and results for the 2022 World Cup are saved in `FIFA_2022_Fixtures_Cleaned.csv`.

9 Challenges

- **Incomplete Data:** Some matches were marked as walkovers, which needed to be removed.
- **Data Formatting:** Ensuring consistent formatting in names and scores was critical to avoid errors during analysis.

10 Future Improvements

- **Incorporating Player Statistics:** Adding player-level data (e.g., goal scorers, assists, and cards) can enhance the accuracy of predictions.
- **Real-time Updates:** Allowing real-time updates based on live match data could further improve predictive performance.

11 Conclusion

This project successfully scrapes, processes, and predicts FIFA World Cup match outcomes using a data-driven approach. By leveraging Poisson regression, it provides a robust model for forecasting match results, which could be applied to future World Cups or other football tournaments.