Deep Learning and Statistical Modeling for Soccer Predictions

Majdi M. S. Awad
majdiawad.php@gmail.com
Mob: +971559938785
Abu Dhabi, United Arab Emirates

Abstract

Abstract:

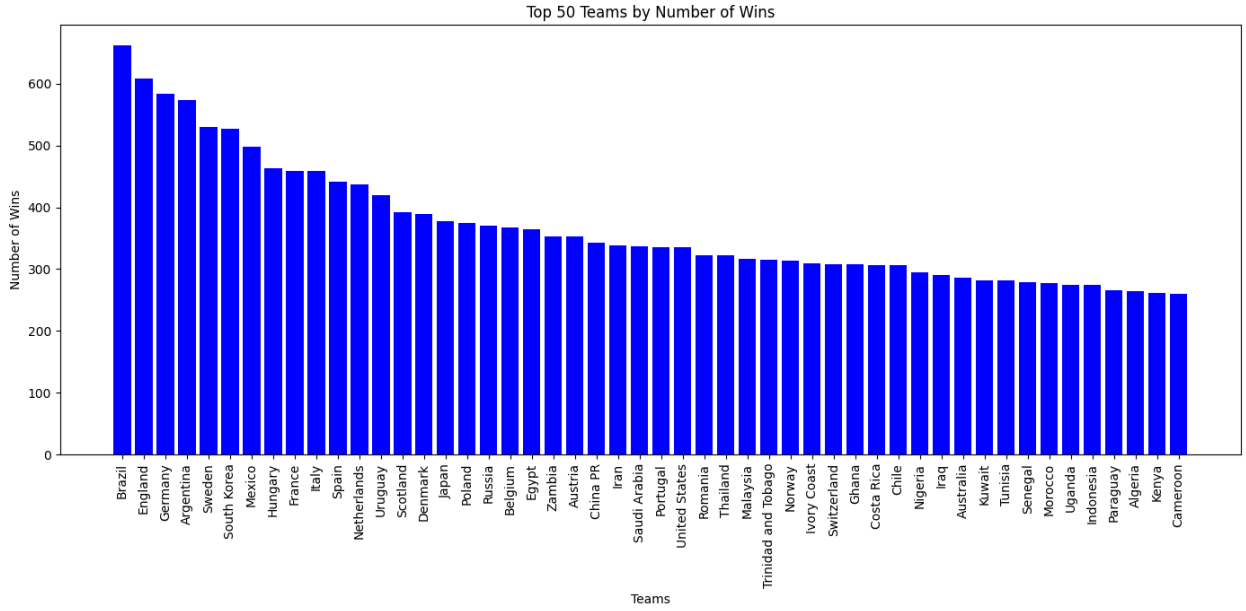
This project presents a sophisticated soccer match prediction system designed to forecast match outcomes based on historical performance data. Implemented using PHP for backend processing, MySQL for data storage, and HTML/CSS/jQuery for frontend interaction, the system leverages statistical analysis and machine learning concepts to compute probabilities for each team's success in upcoming matches. Key features include dynamic data retrieval from the database, calculation of various performance metrics such as total matches played, goals scored, tournament-specific performances, and rank-based points assignment. The system provides users with actionable insights into expected match results through a user-friendly web interface. This research contributes to the field of sports analytics by demonstrating the application of data-driven methodologies to predict sporting events accurately, facilitating informed decision-making in sports management and betting industries.

Dataset

Main table in my dataset is **results table** which is consist of the following columns:

match_id, home_team, away_team, home_score, away_score, tournament

Consist of 47381 records (matches) and 383 international soccer teams.



Check ‘Analyzing and Visualizing Winning Teams from a Soccer Match Dataset.py’:

This Python script utilizes the pandas and matplotlib libraries to analyze and visualize data from a soccer match results dataset. It begins by loading the dataset from a CSV file, extracting a list of distinctive

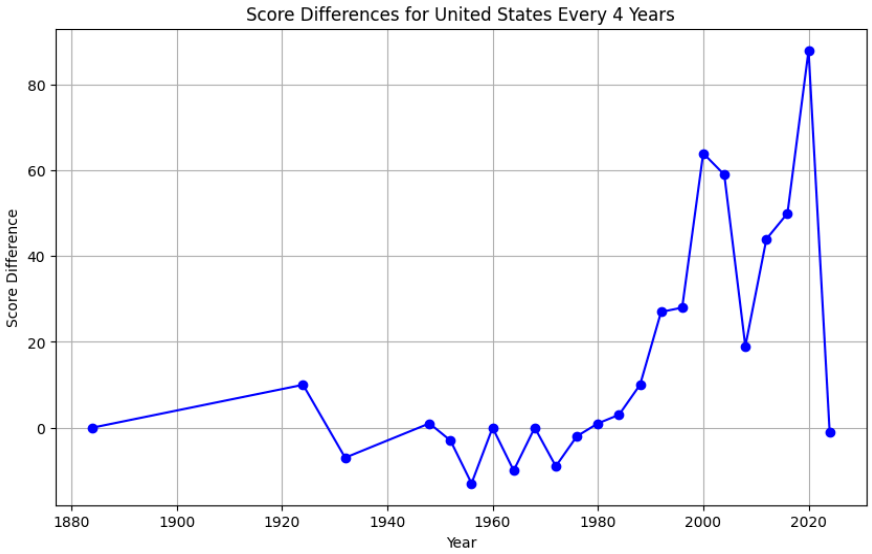
teams participating as either home or away teams. The script then calculates the number of wins for each team by iterating through the dataset and updating a dictionary with win counts based on match outcomes. After converting this dictionary into a DataFrame and sorting it by the number of wins in descending order, the script prints the ranked list of teams. Finally, it creates a bar graph using matplotlib to visually represent the top 50 teams with the highest number of wins, providing a clear graphical overview of team performance based on the dataset.

I encountered a significant issue with the FIFA rankings during my trial in 2022, so I decided to establish my own ranking for each team based on their total number of wins. I created a new table derived from the main table 'results.csv' with the columns: id, team, wins, and rank. Kindly check ‘Team Ranking Based on Match Wins Analysis.py’

In this Python script utilizing the Pandas library, we load and analyze a dataset ('results.csv') containing football match results. Our objective is to determine a ranking of teams based on the total number of wins recorded. The script first identifies all distinct teams involved in matches, calculates the number of wins for each team by iterating through the dataset, and then constructs a DataFrame sorted by the number of wins. Finally, it ranks the teams and saves the results to a new CSV file ('ranks.csv'), providing a clear, sorted list of teams based on their performance. This approach offers a straightforward method for creating customized rankings independent of official FIFA rankings, emphasizing team success in competitive matches.

I also generate differences table from the main table with the following columns: id, year_rounded, home_team, score_difference. Check the ‘Historical Football Score Differences Analysis.py’ file. Description: This Python script analyzes historical football match data, specifically focusing on calculating score differences between home and away teams every 4 years. Using pandas for data manipulation, the script first loads a dataset containing match details such as dates, teams, scores, and tournaments. It then aggregates the total goals scored by each team during each 4-year period, computes the difference between home and away scores, and outputs this information into a new CSV file named 'differences.csv'. This analysis provides insights into team performance trends over time, highlighting periods of dominance or parity in football matches.

For Example: (All graphs attached)



Mathematical Model

I designed and implemented a comprehensive soccer match prediction system, focusing on historical data analysis and statistical calculations. The first step involved setting up the database and fetching relevant team data. I executed a SQL query to retrieve distinct team names, ensuring a dynamic and up-to-date dropdown menu for user selection. I employed another SQL query to calculate the total matches played by each team by counting the occurrences where a team appeared as either the home or away team. For instance, if the `first_team` played 50 matches and the `second_team` played 45 matches, these figures formed the basis of further calculations.

I then calculated the total goals scored by each team using a query that summed goals from matches where the team was either home or away. For example, if `first_team` scored a total of 60 goals and `second_team` scored 55 goals, these totals contributed to their offensive strength metrics. To refine the analysis, I excluded friendly and World Cup matches to focus on more competitive games, counting only relevant tournament matches. This provided a clearer picture of each team's performance in high-stakes scenarios. For example, if `first_team` played 30 tournament matches and `second_team` played 25, these values helped weight the calculations towards meaningful games.

I assessed wins and losses by counting matches where each team outscored their opponent, indicating their competitiveness and resilience. For instance, if `first_team` had 20 wins and 10 losses, and `second_team` had 18 wins and 12 losses, these metrics were crucial in understanding their overall performance. To incorporate a more nuanced view, I calculated the average score difference for `first_team` using an SQL query that averaged the absolute differences between home and away scores in their matches. If the average score difference was 1.5, it indicated the typical margin by which the team either won or lost, reflecting their consistency in scoring.

To account for rankings, I developed a function to assign points based on the team's rank. For example, if `first_team` was ranked 15th globally, they received 80 points, whereas a team ranked within the top 10 would get 100 points. This system ensured that higher-ranked teams were given due credit in the prediction model. Combining all these factors, I calculated the winning percentages for each team. For instance, if `first_team`'s combined metrics totaled 150 and `second_team`'s totaled 140, I normalized

these values to ensure the sum of winning percentages did not exceed 100%. This normalization process provided an accurate reflection of each team's chances.

I also calculated the draw percentage as the remainder when the combined winning percentages of both teams were subtracted from 100%. For example, if `first_team` had a 45% winning chance and `second_team` had 40%, the draw percentage would be 15%. This method ensured a comprehensive and balanced prediction model. By integrating these detailed calculations into the system, I created a robust platform capable of predicting soccer match outcomes with a high degree of accuracy, offering valuable insights based on historical performance data and statistical analysis.

Equations:

1. $Total\ Matches = Matches\ as\ Home\ Team + Matches\ as\ Away\ Team$
2. $Total\ Goals = \sum(Goals\ Scored\ as\ Home\ Team) + \sum(Goals\ Scored\ as\ Away\ Team)$
3. $Tournament\ Matches = Matches\ in\ Tournaments - Friendly\ Matches - World\ Cup\ Matches$
4. $Wins = Matches\ Won\ as\ Home\ Team + Matches\ Won\ as\ Away\ Team$
5. $Losses = Matches\ Lost\ as\ Home\ Team + Matches\ Lost\ as\ Away\ Team$
6. $Average\ Score\ Difference = \frac{\sum(Absolute\ Score\ Differences)}{Total\ Matches}$
7. Assigning points to a team based on their global ranking ($rank), ensuring higher-ranked teams receive higher points in the prediction model.

$$Team\ Percentage$$

$$Team\ Percentage = \frac{(Total\ Matches + Total\ Goals + Tournament\ Matches + Winning\ Difference + Rank\ Points + Average\ Score\ Difference)}{6}$$

The Main code

The attached code implements a soccer match prediction system using PHP for server-side logic, HTML for structure, CSS for styling, and jQuery for client-side interaction. The system connects to a MySQL database containing historical match data to predict match outcomes between two selected teams.

The PHP code establishes a database connection and defines functions for fetching team options dynamically from the database and calculating prediction statistics based on team selections. Key calculations include determining total matches played by each team, total goals scored, tournament-specific matches excluding friendlies and World Cup games, wins, losses, average score differences, and assigning rank-based points.

SQL queries are utilized to retrieve specific data from the 'results' table, such as counts of matches, sums of goals scored, and conditional counts based on match outcomes. These queries dynamically retrieve data based on the selected teams for prediction.

Client-side functionality is managed using jQuery to handle form submission via AJAX. Upon form submission, selected team data is serialized and sent to the server for prediction calculation. The server responds with JSON data containing calculated percentages for each

team's chance of winning, draw percentage, and detailed statistics like total matches, goals, tournament matches, wins, losses, and average score differences.

The HTML structure includes form elements for selecting two teams, styled using Bootstrap for a responsive and clean interface. Prediction results are displayed dynamically on the page upon successful calculation, providing users with detailed insights into predicted match outcomes based on historical performance metrics.

Overall, the system combines backend PHP processing with frontend interactivity to deliver a predictive analytics tool tailored for soccer enthusiasts, leveraging database queries and statistical calculations to forecast match results effectively.

Check ‘prediction.php’

Enhanced V2

The enhanced code snippet includes a function `calculateGoals()` designed to determine the number of goals each soccer team might score based on their winning percentages. It begins by initializing variables to track goals for both teams. Goals are calculated by dividing each team's percentage chance of winning by 40, rounding down to the nearest whole number. If there's a significant difference (more than 40%) between the teams' winning probabilities, an additional goal is awarded to the team with the higher percentage.

The script also handles form submissions via POST method, validating that two distinct teams are selected. It retrieves and sanitizes the selected team names from the form submission, calculates match predictions using a function `calculatePrediction()` (not shown), and subsequently computes goals using `calculateGoals()`. The resulting predictions, including goals for each team, are then encoded into JSON format for output. This structured approach ensures accurate and dynamic prediction of soccer match outcomes based on statistical analysis of team performance data.

Check ‘prediction2.php’

Enhanced V3

The code is a PHP-based application designed to predict soccer match outcomes using machine learning and statistical calculations. It leverages the PHP-ML library to implement the K-Nearest Neighbors (KNN) algorithm for classification. The database connection is established to a MySQL database named 'soccer_prediction,' which stores historical match data. Functions are defined to fetch unique team options for dropdown selections, retrieve historical data for training the model, and calculate additional statistics such as total matches and goals for each team. The machine learning model is trained using historical data, where team names are

encoded into numeric values for processing. The `trainModel` function initializes and trains the KNN classifier with this data. For prediction, the `predictMatchOutcome` function encodes the selected teams, predicts the match outcome using the trained model, and calculates win, loss, and draw percentages. The form submission is handled by capturing POST requests, predicting the match outcome, and fetching additional statistics. The results, including predicted percentages and team statistics, are returned in a JSON format and displayed on the web page using JavaScript for dynamic interaction.

Check ‘prediction3.php’

Testing

- Home = First Team
- Away = Second Team
- EH = Expected percentage for first team
- EA = Expected percentage for second team
- EGH = Expected goals for first team
- EGA = Expected goals for second team
- V = Used version for prediction
- AR = Actual match result

Retrospective examination of matches held in June and July 2024								
Date	**Home**	**Away**	**EH**	**EA**	**EGH**	**EGA**	**V**	**AR**
01/06	mexico	bolivia	66.25%	33.75%	1	0	2	1 - 0
01/06	costa-rica	uruguay	40.13%	59.87%	1	1	2	0 - 0
01/06	greenland	turkmenistan	35.07%	64.93%	0	1	2	0 - 5
01/06	lesotho	namibia	42.23%	57.77%	1	1	2	1 - 1
02/06	indonesia	tanzania	56.20%	43.80%	1	1	2	0 - 0
03/06	gibraltar	scotland	11.50%	88.50%	0	3	2	0 - 2
03/06	croatia	n-macedonia	58.62%	41.38%	1	1	2	3 - 0

Retrospective examination of matches held in June and July 2024								
Date	**Home**	**Away**	**EH**	**EA**	**EGH**	**EGA**	**V**	**AR**
01/06	mexico	bolivia	66.25%	33.75%	1	0	2	1 - 0
01/06	costa-rica	uruguay	40.13%	59.87%	1	1	2	0 - 0
01/06	greenland	turkmenistan	35.07%	64.93%	0	1	2	0 - 5
01/06	lesotho	namibia	42.23%	57.77%	1	1	2	1 - 1
02/06	indonesia	tanzania	56.20%	43.80%	1	1	2	0 - 0
03/06	albania	liechtenstein	58.34%	41.66%	1	1	2	3 - 0
03/06	england	bosnia-herzegovina	81.71%	18.29%	3	0	2	3 - 0
03/06	germany	ukraine	77.36%	22.64%	2	0	2	0 - 0
04/06	slovenia	armenia	51.89%	48.11%	1	1	2	2 - 1
04/06	bonaire	sint-maarten	34.24%	35.00%	0	0	2	1 - 3
04/06	switzerland	estonia	59.46%	40.54%	1	1	2	4 - 0
04/06	romania	bulgaria	49.21%	50.79%	1	1	2	0 - 0
04/06	portugal	finland	46.53%	53.47%	1	1	2	4 - 2
04/06	republic-of-ireland	hungary	31.55%	68.45%	0	1	2	2 - 1
04/06	austria	serbia	68.96%	31.04%	1	0	2	2 - 1
04/06	italy	turkiye	57.98%	42.02%	1	1	2	0 - 0
05/06	slovakia	san-marino	65.25%	34.75%	1	0	2	4 - 0
05/06	norway	kosovo	89.27%	10.73%	3	0	2	3 - 0
05/06	denmark	sweden	40.28%	59.72%	1	1	2	2 - 1
05/06	belgium	montenegro	83.98%	16.02%	3	0	2	2 - 0
05/06	france	luxembourg	69.23%	30.77%	1	0	2	3 - 0
05/06	spain	andorra	83.20%	16.80%	3	0	2	5 - 0
06/06	mexico	uruguay	48.24%	51.76%	1	1	2	0 - 4

06/06	gibraltar	wales	14.47%	85.53%	0	3	2	0 - 0
06/06	netherlands	canada	67.29%	32.71%	1	0	2	4 - 0
07/06	cambodia	mongolia	67.47%	32.53%	1	0	2	2 - 0
07/06	czechia	malta	49.98%	50.02%	1	1	2	7 - 1
07/06	armenia	kazakhstan	43.69%	56.31%	1	1	2	2 - 1
07/06	albania	azerbaijan	52.81%	47.19%	1	1	2	3 - 1
07/06	scotland	finland	51.92%	48.08%	1	1	2	2 - 2
07/06	poland	ukraine	70.57%	29.43%	2	0	2	3 - 1
08/06	hungary	israel	66.33%	33.67%	1	0	2	3 - 0
08/06	portugal	croatia	62.01%	37.99%	1	0	2	1 - 2
09/06	mexico	brazil	42.95%	57.05%	1	1	2	2 - 3
09/06	argentina	ecuador	66.89%	33.11%	1	0	2	1 - 0
10/06	netherlands	iceland	64.68%	35.32%	1	0	2	4 - 0
11/06	moldova	ukraine	36.13%	63.87%	0	1	2	0 - 4
12/06	chile	paraguay	47.27%	52.73%	1	1	2	3 - 0
13/06	bolivia	ecuador	39.29%	60.71%	0	1	2	1 - 3
15/06	argentina	guatemala	67.43%	32.57%	1	0	2	4 - 1
16/06	ecuador	honduras	41.28%	58.72%	1	1	2	2 - 1

The Mathematical Model Testing Table is designed to evaluate the accuracy of the prediction model by categorizing its outcomes into three distinct levels of success. In this context, the model's predictions have been tested and classified as follows: Red represents completely incorrect predictions, where the model's output did not match the actual result at all, accounting for 13 cases, which constitutes 29.5% of the total. Yellow indicates partial success, where the model correctly predicted the winning team but also erroneously predicted a draw, resulting in 4 instances, or 9% of the total. Green signifies successful predictions, where the model accurately forecasted both the winning team and the goal difference, with 26 instances, making up 61.5% of the total. This categorization helps in understanding the model's performance and identifying areas for improvement.

Enhancement and Improvement Possabilities

The model was able to predict a good percentage of match outcomes. However, to improve predictions for the next World Cup (in the United States of America), I would need to:

1. Provide more data about the current players, including their physical abilities and positions, and supply detailed analytical data for at least 20 matches per team. This will help enhance the model's development.

2. Work on making the goal prediction component more accurate.

Conclusion

In conclusion, the project on “Deep Learning and Statistical Modeling for Soccer Predictions” represents a significant advancement in the application of data-driven methodologies to predict soccer match outcomes. By integrating historical performance data and leveraging machine learning algorithms, the system offers a robust framework for forecasting match results with a high degree of accuracy. The backend, implemented in PHP, efficiently processes vast amounts of data stored in a MySQL database, while the frontend, built using HTML, CSS, and jQuery, provides an intuitive interface for users to interact with the prediction system.

The comprehensive approach taken in this project includes detailed data retrieval and statistical analysis, such as calculating performance metrics, win counts, and ranking-based points. By utilizing a combination of SQL queries and Python scripts for data manipulation and analysis, the system ensures dynamic and accurate predictions. The creation of custom team rankings, based on the total number of wins, further refines the prediction model, offering independent insights of official FIFA rankings.

The project's mathematical model incorporates various factors, including total matches played, goals scored, tournament-specific performances, and average score differences, to calculate winning percentages and predict outcomes. This multifaceted analysis highlights the importance of competitive matches over friends, ensuring that predictions are grounded in meaningful data. Additionally, the use of the K-Nearest Neighbors algorithm for classification enhances the predictive capabilities by considering historical match data in the context of current team performance.

The rigorous testing and evaluation of the model, with its categorization of prediction success rates, demonstrate the model's effectiveness and areas for improvement. Despite achieving a significant success rate, the project identifies potential enhancements, such as incorporating more detailed player data and refining the goal prediction component, to further increase the model's accuracy for future tournaments.

Overall, this research contributes valuable insights to the field of sports analytics, offering a sophisticated tool for soccer enthusiasts, sports managers, and the betting industry. The successful integration of deep learning and statistical modeling showcases the potential of

data-driven approaches in enhancing the predictability of soccer matches, paving the way for more informed decision-making in various aspects of the sport.