Predicting the Euro 2024 Bracket

Ben Ghouzi, Anzar Chowdhury, Triya Basu & Emma Rabbath

ABSTRACT

Sports analytics has grown significantly over the past decade, impacting the way teams prepare for matches, scouts evaluate players, and how fans engage with sports. Football, the world's most popular sport, has seen a considerable application of data science techniques (to predict..., to evaluate?). Euro 2024 is a unique opportunity to explore the unpredictability of football matches, where outcomes can be influenced by a variety of factors such as team form, player injuries, weather, location, and refereeing decisions. This project aims to explore the analytics of predicting match outcomes, offering insights that could benefit both teams and fans.

INTRO

Defining the Problem

Predicting the outcomes of football matches is a very challenging task as multiple factors are in play. It goes beyond analyzing statistical data, team form, and historical performances. The unpredictable nature of the sport has led to unexpected results that differ from expert analysis which we have seen throughout the history of the game.

Motivation for the Topic

Data analytics is becoming more and more important in sports. Football, being the most watched sport globally, offers a vast and dynamic dataset encompassing variables influencing match outcomes. We aim to contribute to the ongoing discourse surrounding sports analytics by diving into the analytics of predicting football match outcomes. We are addressing the needs of teams, fans, and betting markets. Accurate match predictions hold the potential to profoundly impact team strategies, aiding in opponent analysis, player selection, and tactical decisions.
Furthermore, they foster heightened engagement and excitement among fans and betting markets, enriching the overall spectator experience. Additionally, the timing of Euro 2024 makes it a one-of-a-kind chance to study how football games change during an international event. The event has teams from all over Europe playing at the highest level. This means that there are a lot of different matchups, playing styles, and factors that can affect the outcome that can be studied and added to our prediction model. Our motivation for embarking on this project lies at the intersection of sports, data science, and the desire to understand the unpredictable nature of football matches.

RELATED WORK

One notable project which is fairly similar to our goal would be Predicting FIFA 2022 World Cup with ML. This project aims to predict the results of the FIFA World Cup using Machine Learning, this project is also using the same dataset as us. A second related project is Predicting Football Match Outcomes with Machine Learning. This projects highlights the difficulties of predicting outcomes in a sport as unpredictable as football. However, the author does conclude by saying that “while no model can guarantee 100% accuracy due to the inherent unpredictability of the sport, machine learning can provide valuable insights and increase the likelihood of making informed predictions.”

METHODOLOGY

Data Acquisition

We found two datasets from Kaggle for our Euro 2024 Bracket Model. Analyzing previous trends, match records and team rankings will allow us to make accurate predictions for the future.

Data Source 1: International Football Results from 1872 to 2024

This dataset contains international football results starting from the first official mach in 1872 to today. It includes information such as match date, team playings, score, type of tournament, city, and country. The dataset is useful for analyzing trends in international football, comparing team performances, and more.

FIFA World Ranking 1992 - 2023

This dataset contains the FIFA World Rankings for men’s national footbal teams. The rankings are based on team’s game results, with more recent results and more significant matches have a larger weight. This dataset is useful for us as our goal would be to predict the EURO 2024 bracket.

DATA PREPARATION STEPS

In the initial stages of our exploratory data analysis (EDA) for FIFA match results and rankings, we began by importing the dataset covering matches from 1872 to 2024 using the pandas library in Python. To set up our analysis, we first standardized the date formats and then focused on matches post-January 2014 due to previous criticisms of the ranking system. We addressed inconsistencies in country names, such as 'Czech Republic' being listed as 'Czechia', and standardized data for accurate comparisons. After filtering to include only UEFA teams and merging datasets of match results with FIFA rankings, we sorted the data chronologically to avoid bias. We also added a column to indicate whether the home team was ranked higher than the away team, enhancing our analysis of match outcomes based on team rankings. This process of cleaning up the data ensured a reliable dataset for further analysis into team performance trends and ranking dynamics.

MODEL SELECTION

For our project, we've selected to employ three algorithms: Random Forest Classifier, KNeighbors Classifier, and Support Vector Machine (SVM).

Random Forest Classifier

This model incorporates a variety of features, including the home and away team rankings, changes in rankings, and match outcomes. We've created features such as 'form points', which represents the total points a team has earned in their last five matches, and 'head-to-head performance', which represents the result of the last encounter between two teams (win, loss, or draw). Additionally, we've created a 'points change' feature to track the difference in team rankings since FIFA's last publication. To evaluate the accuracy of our model we used accuracy score and a classification report which gives precision, recall, and F1-score.

KNeighbors Classifier

For this second model, we used the StandardScaler to normalize our data, an important step for distance-based algorights to work correctly. We trained the model on the scaled dataset and assessed its performance using a classification report.

SVM

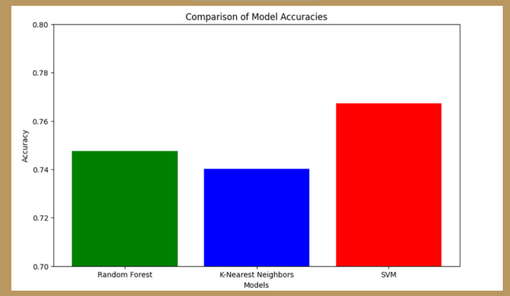
For our third model, we used the StandardScaler to normalize our data, crucial for the RBF kernel's performance. We once again trained the model on the scaled dataset and assessed its performance using a classification report.

The model's efficacy is determined through various metrics, notably the accuracy score, which offers a quick overview of the performance by measuring the number of correct predictions across the dataset. This is useful in our context of predicting UEFA Euro 2024 outcomes, providing a straightforward measure of the model's reliability.

RESULTS AND ANALYSIS

Final Bracket

In general, our goals for this project were met because we were able to predict the winner of the Euro 2024 based on our model. In the Euro 2024 tournament, England was the winner according to our model. The match came down to a game between Germany vs England and Belgium vs Italy. According to those results, England and Italy had won those respective games. When it came down to the final, England won against Italy. According to the results of the Euro 2020, Italy had won against England. The consistency between our model's final two teams, Italy and England, and the actual outcome of the previous Euro Cup adds credibility to the accuracy and reliability of our prediction model. Overall, our model's ability to predict the Euro 2024 winner and showcase historical outcomes shows its effectiveness in analyzing football tournaments and predicting match outcomes.



SVM shows the highest accuracy among the three models, which suggests it might be the best at handling the complexity of the dataset with its ability to model non-linear decision boundaries effectively. Random Forest also performs well. This model benefits from its ability to handle high-dimensional data and model non-linear relationships. K-Nearest Neighbors has a slightly lower accuracy compared to the other two.
The SVM model shows the highest overall accuracy among the three models. An accuracy of 0.77 is indicative of a model that performs well across all outcome classes. The macro and weighted averages indicate that the SVM model treats all classes fairly and performs well even when class imbalance is accounted for. This suggests that the model is not only performing well on the majority class but also handling minority classes effectively. The SVM's ability to provide high precision and recall, particularly for home wins, which are more frequent in football, demonstrates its robustness as a model. The consistent performance across different classes suggests that the SVM model has a good generalization capability, which is crucial for the unpredictable nature of football matches.

IMPACTS

In this project, it's evident that the application of machine learning to predict football match outcomes, including specifically predicting the winner, has significant impacts across the sports industry. By increasing prediction accuracy, our project helps teams in strategic planning, player selection, and accordingly making adjustments. This increases fan engagement by giving us more analytical insights. It benefits the betting community by giving them more accurate predictions, leading to better bets and increased engagement. Predictions of match outcomes enhance marketing strategies, optimize ticket sales, and attract sponsorships through targeted advertising during key matches. This project also shows the importance of data-driven decision-making processes, showing how previous records can help us make optimal future decisions. Overall, the project not only advances predictive capabilities in football but also promotes a broader appreciation and strategic use of data across various industries.

CONCLUSION

For our project, we used these three models: SVM, KNN, and Random Forest.
The SVM model stood out, achieving an overall accuracy of 77%. It demonstrated robustness in handling both the majority and minority classes effectively. The model's precision and recall were particularly impressive for predicting home wins, which are more frequent in football. The KNN and Random Forest models also contributed valuable insights, with accuracies between 70% to 74%. The models performed well in predicting home wins. This is likely because of the fact that teams playing on their home field often have better outcomes. However, our model struggled to accurately predict draws because the outcome is more complex and less predictable.
Going forward, we could further improve by incorporating more data such as in-game events and maybe even psychological factors. We can explore team dynamics, which can help us understand more about a player's motivation. Since this has an impact on their overall performance, it is definitely important to consider.

References

International Football Results 1872-2024

IFA World Ranking 1992 - 2023raph text