Report

Shuhao Liu, Jianpeng Chen, Chenyu Song, Heshun Wang

A. Abstract

In this project, we aim to predict the outcomes of UEFA 2024 using data from international football matches between 1872 and 2017, and current FIFA rankings. The European Cup, starting in 1955, has grown into a major global football event, making accurate predictions valuable for team strategies, the betting industry, and market strategies for companies and sponsors.

The application of machine learning in predicting football results has proven effective. Our methodology involves using advanced machine learning techniques such as logistic regression, decision trees, random forests, support vector machines, neural networks, and particularly ensemble methods like gradient boosting and XGBoost. These methods are chosen for their ability to improve accuracy by analyzing patterns in large historical data sets of 46,442 international match results, including diverse competitions and friendly matches.

Our preliminary tests, including the 2022 World Cup predictions, have shown promising accuracy, especially in predicting the top 16 and top 8 outcomes. We anticipate applying these techniques to Euro 2024 will enhance prediction accuracy, supporting strategic decisions in sports management, betting industries, and economic planning.

By leveraging these advanced analytics, we aim to contribute to the evolution of sports analytics, providing crucial insights that benefit various stakeholders and potentially influencing future trends in sports predictions as data and technology continue to advance.

B. Introduction

In this project, we aim to predict the outcomes of the UEFA European Championship starting in June 2024 by analyzing historical international football competition data from 1872 to 2017 alongside current FIFA rankings. The UEFA European Championship, commonly referred to as the Euros, has been one of the most highly anticipated football events globally since its inception in 1955. The tournament brings together Europe's elite national teams to compete for the continent's top football honor.

Predicting the results of such a highly watched event involves complex data analysis, given the unpredictability of football matches. However, advancements in technology and machine learning algorithms have now made it possible to achieve high-accuracy predictions. These predictions are of significant value to multiple stakeholders:

- Impact on Teams: Coaches and team managers can utilize predictive insights to tailor their tactical and training focuses, potentially enhancing team performance during the tournament.
- Impact on the Betting Industry: Football betting constitutes a substantial market, and accurate match predictions are crucial for both bookmakers and bettors. Predictive results can influence odds setting and betting strategies, affecting how bets are placed.
- Impact on Business and Marketing Decisions: Companies integrate their marketing efforts with major sporting events through sponsorships and related activities. Predicting match outcomes can help businesses formulate more effective marketing strategies and investment decisions.

Importance of Solving This Problem:

Understanding past performance of teams in games is crucial for predicting their future performance. By analyzing historical data, patterns of wins, losses, scoring habits, and performance under specific conditions such as home or away games, or against certain opponents can be identified. These statistics are instrumental in assessing the overall strength of a team.

Literature Survey:

Research on predicting outcomes of the Euros has evolved with the application of machine learning technologies. Common methodologies include logistic regression, random forests, and neural networks, which manage large datasets and identify complex patterns. Jawade et al. analyzed various machine learning models used for predicting football match outcomes, providing a comprehensive overview of the field and identifying current research trends and gaps[1]. Penn et al. utilized a bivariate Poisson model to predict the outcomes of the 2020 European Championship[2], demonstrating the effectiveness of sophisticated statistical models in sports predictions. Ren et al. and Carloni et al. proposed deep learning frameworks that use neural networks to analyze historical data and team performance metrics for predicting football match results[3][4]. These studies not only demonstrate the effectiveness of different models and methods but also provide insights into the future direction of football match prediction technologies.

Predicting the outcomes of the Euros is not just a showcase of sports analytics technology but also involves significant economic, media, and cultural implications. By applying advanced data analysis and machine learning techniques, the accuracy of predictions can be enhanced, making the process increasingly complex and scientific. This project exemplifies how technological progression is continuously reshaping the landscape of sports predictions, impacting various aspects of the sports industry and beyond. Through this endeavor, we aim to provide a comprehensive analysis that benefits teams, bettors, and commercial enterprises associated with the event, offering insights that could potentially transform future trends in sports predictions as data and technology continue to advance.

C. Methodology

In our project, we use statistical machine learning methods to predict the outcome of football matches. The main advantage of machine learning is its ability to automatically process and analyze large data sets, learning from complex data and identifying patterns and trends that can help with predictions. This method is more efficient and precise than traditional manual analysis because it can use algorithms to discover subtle correlations in the data that are not easy to detect.

The historical football match data used in our project provides a solid foundation for the model. The data contains rich information on match results, team performance, match location and match history. With this data, we can build and train models to predict future match outcomes. In the selection and use of models, we use a variety of machine learning technologies to ensure that data can be analyzed from multiple angles and improve the accuracy of predictions.

Specific methods include:

- Logistic regression: This is a statistical model suitable for binary classification problems and is often used to predict the outcome of a game. Logistic regression can evaluate the impact of different factors (such as team status, historical winning percentage, etc.) on the outcome of the game.
- Decision Trees and Random Forests: Decision trees predict outcomes by creating a model that splits a data set into different branches. Random forests are composed of multiple decision trees, which as a whole provide more stable and accurate predictions.
- Neural Networks: This powerful model is capable of capturing non-linear relationships in data and is particularly suitable for processing large-scale data sets, and its predictive capabilities can be further improved through deep learning.
- Ensemble methods: Such as gradient boosting machine (GBM) and XGBoost, these methods can significantly improve the accuracy of predictions by combining multiple weaker models to form a powerful model.

Through these advanced machine learning methods, we can continuously optimize the model and improve the accuracy of predicting the results of future football matches. This has great significance for game strategy development, betting industry and related business decisions.

D. Code

In our project, we divide the code into two core parts: data processing and exploratory data analysis (EDA), and model training, prediction, and validation.

1. Data processing and EDA:

The first step in data processing is data cleaning. We deal with missing values in the data set to ensure the integrity and accuracy of the data. After cleaning, we perform exploratory data analysis, a process that involves conducting in-depth analysis of the data set to identify patterns, trends, and relationships in the data. For example, we analyzed the changes in FIFA rankings of each national team in different years, the frequency of various types of competitions, and goal statistics. Through these analyses, we are able to gain insights into team performance and match characteristics that are critical for feature selection and subsequent model training.

2. Training model, prediction and verification:
In the feature selection stage, we select those features that are highly relevant to the game results based on the findings of the EDA stage. Next, we apply multiple machine learning algorithms to train the predictive model. We tried logistic regression, decision trees, random forests, neural networks, and ensemble methods such as gradient boosting machines (GBM) and XGBoost. Each of these models has its own advantages, such as logistic regression having advantages in interpreting results, while ensemble methods show high accuracy when dealing with complex data structures. By comparing and analyzing the training results of different models, we finally selected the logistic regression model as the main tool for predicting the results of the 2024 European Cup.

In order to verify the accuracy and practicality of the model, we used historical data for backtesting. In particular, we used the model to predict the results of the 2022 World Cup and compared these predictions with the actual results. This verification method not only helps us evaluate the performance of the model in practical applications, but also provides a direction for further optimization of the model. In this way, we were able to verify the validity of the model and be confident that our model can provide reliable predictions for the upcoming European Cup.

In short, through systematic data processing and accurate model training, as well as strict model verification, our project aims to provide a scientific and practical tool to predict the results of football matches, which not only has guiding significance for team strategy formulation, but also Also provides support for the gaming industry and related business decisions.

E. Results

First, obtain the results of international football matches from 1872 to 2024. The data consists of (date, home_team, away_team, home_score, away_score, tournament, city, country, neutral). As shown below.

	date	home_team	away_team	home_score	away_score	tournament	city	country	neutral
0	1872-11-30	Scotland	England	0	0	Friendly	Glasgow	Scotland	False
1	1873-03-08	England	Scotland	4	2	Friendly	London	England	False
2	1874-03-07	Scotland	England	2	1	Friendly	Glasgow	Scotland	False
3	1875-03-06	England	Scotland	2	2	Friendly	London	England	False
4	1876-03-04	Scotland	England	3	0	Friendly	Glasgow	Scotland	False

Then obtain the world rankings of all national teams from December 31, 1992 to July 20, 2023. The results for the first five rows are shown below.

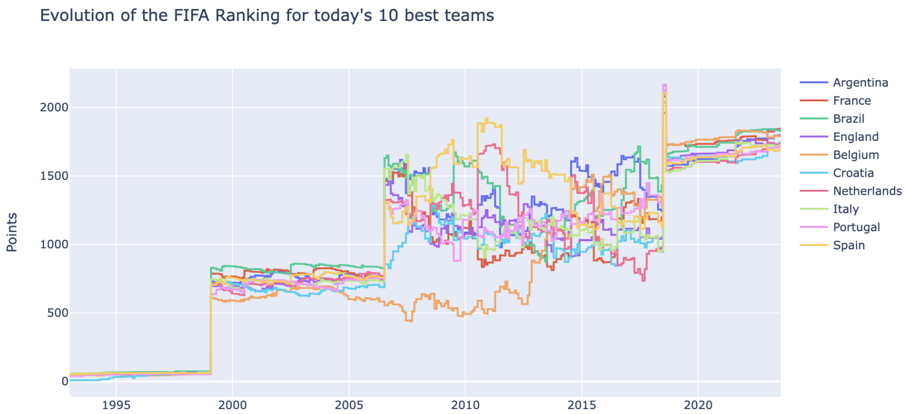
	rank	country_full	country_abrv	total_points	previous_points	rank_change	confederation	rank_date
0	1	Germany	GER	57.0	0.0	0	UEFA	1992-12-31
1	96	Syria	SYR	11.0	0.0	0	AFC	1992-12-31
2	97	Burkina Faso	BFA	11.0	0.0	0	CAF	1992-12-31
3	99	Latvia	LVA	10.0	0.0	0	UEFA	1992-12-31
4	100	Burundi	BDI	10.0	0.0	0	CAF	1992-12-31

Then through data sorting, the scores of the same country in each country ranking are sorted out. The chart below is an example of scores for the top five countries.

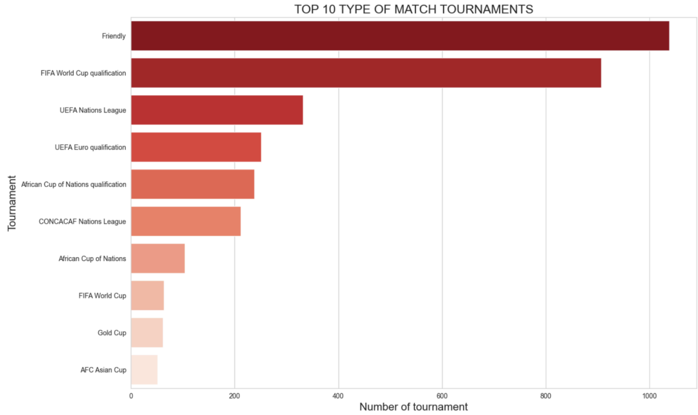
rank_date	1992-12-31	1993-01-01	1993-01-02	1993-01-03	1993-01-04	1993-01-05	1993-01-06	1993-01-07	1993-01-08	1993-01-09	...	2023-07-11	2023-07-12	2023-07-13	2023-07-14	2023-07-15	2023-07-16	2023-07-17	2023-07-18
country_full																			
Afghanistan	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1020.32	1020.32	1020.32	1020.32	1020.32	1020.32	1020.32	1020.32
Albania	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	14.0	...	1357.39	1357.39	1357.39	1357.39	1357.39	1357.39	1357.39	1357.39
Algeria	39.0	39.0	39.0	39.0	39.0	39.0	39.0	39.0	39.0	39.0	...	1511.23	1511.23	1511.23	1511.23	1511.23	1511.23	1511.23	1511.23
American Samoa	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Andorra	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1022.30	1022.30	1022.30	1022.30	1022.30	1022.30	1022.30	1022.30

5 rows × 11159 columns

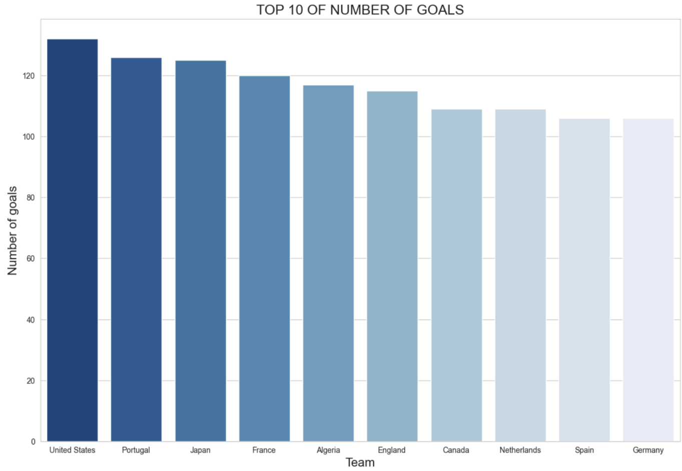
From this we can get a chart of the ranking changes of the ten best national teams in the world today.



The picture below is the statistics of the categories of games we obtained. It can be clearly seen that in addition to the friendly matches with the most games, the most data is from the FIFA World Cup qualifying matches. In the process of model training, we should pay attention to the type of predicted games, which is also an important basis for the success of data mining.



The figure below shows the statistics of the number of goals scored by each national team in the national team competition (Top 10).



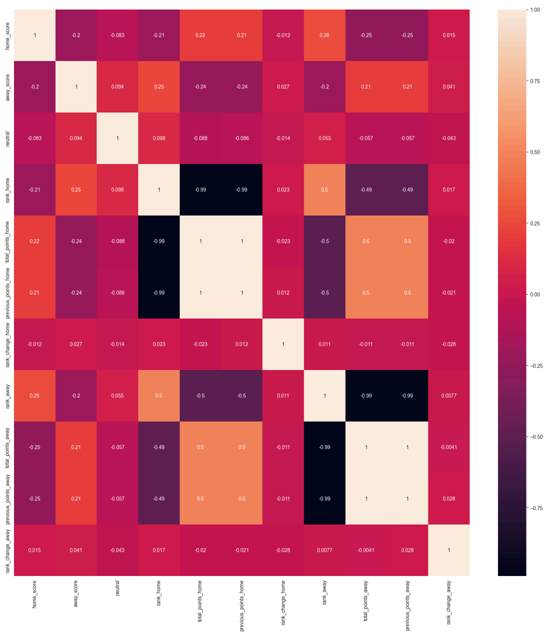
The picture below shows us obtaining the schedule information of each team according to the actual grouping situation and schedule of the 2024 European Cup.

	Group	Team	Country_Name_Short	First match against	Second match against	Third match against
0	A	Germany	GER	Scotland	Hungary	Switzerland
1	A	Hungary	HUN	Switzerland	Germany	Scotland
2	A	Scotland	SCO	Germany	Switzerland	Hungary
3	A	Switzerland	SUI	Hungary	Scotland	Germany
4	B	Spain	ESP	Albania	Croatia	Italy
5	B	Albania	ALB	Spain	Italy	Croatia
6	B	Croatia	CRO	Italy	Spain	Albania
7	B	Italy	ITA	Croatia	Albania	Spain
8	C	England	ENG	Denmark	Slovenia	Serbia
9	C	Denmark	DEN	England	Serbia	Slovenia
10	C	Slovenia	SLO	Serbia	England	Denmark
11	C	Serbia	SRB	Slovenia	Denmark	England
12	D	France	FRA	Austria	Netherlands	Poland
13	D	Austria	AUT	France	Poland	Netherlands
14	D	Netherlands	NED	Poland	France	Austria
15	D	Poland	POL	Netherlands	Austria	France
16	E	Belgium	BEL	Romania	Slovakia	Ukraine
17	E	Romania	ROU	Belgium	Ukraine	Slovakia
18	E	Slovakia	SVK	Ukraine	Belgium	Romania
19	E	Ukraine	UKR	Slovakia	Romania	Belgium
20	F	Portugal	POR	Turkey	Czech Republic	Georgia
21	F	Turkey	TUR	Portugal	Georgia	Czech Republic
22	F	Czech Republic	CZE	Georgia	Portugal	Turkey
23	F	Georgia	GEO	Czech Republic	Turkey	Portugal

We then integrated the results of each match with the national team's world ranking points (ranking points) on that match date to input more dimensions into our dataset.

	date	home_team	away_team	home_score	away_score	tournament	city	country	neutral	rank_date_home	...	rank_change_home	confederation_home	rank_date_away	rank_away	country_full_away	country_abrv_away	total_points_away	previous_points_away	rank_change_away
0	2020-01-07	Barbados	Canada	1	4	Friendly	Irvine	United States	True	2020-01-07	...	0.0	CONCACAF	2020-01-07	73.0	Canada	CAN	1331.0	1331.0	0.0
1	2020-01-09	Moldova	Sweden	0	1	Friendly	Doha	Qatar	True	2020-01-09	...	0.0	UEFA	2020-01-09	17.0	Sweden	SWE	1579.0	1579.0	0.0
2	2020-01-10	Barbados	Canada	1	4	Friendly	Irvine	United States	True	2020-01-10	...	0.0	CONCACAF	2020-01-10	73.0	Canada	CAN	1331.0	1331.0	0.0
3	2020-01-12	Kosovo	Sweden	0	1	Friendly	Doha	Qatar	True	2020-01-12	...	0.0	UEFA	2020-01-12	17.0	Sweden	SWE	1579.0	1579.0	0.0
4	2020-01-15	Canada	Iceland	0	1	Friendly	Irvine	United States	True	2020-01-15	...	0.0	CONCACAF	2020-01-15	39.0	Iceland	ISL	1464.0	1464.0	0.0

Then we use the corr() method of pandas DataFrame to calculate the pairwise correlation of all numeric columns in df. We calculated the Pearson correlation coefficient, which has values between -1 (perfect negative correlation) and +1 (perfect positive correlation), with 0 indicating no correlation. and visualize the results.



Through this result, relevant data features are found and the features are integrated (dimension adjustment). For example, rank_away and rank_home are merged, and rank_away-rank_home is used to replace the above two features.

We train various models (logistic regression; decision trees and random forests; neural networks; ensemble methods such as gradient boosting machines (GBM) and XGBoost. These high-performance algorithms can improve the accuracy of predictions by building multiple models)), predicting the results of the 2024 European Cup group stage.

The results of the linear regression model are shown below.

```
___Starting group C:___
England vs. Denmark: England wins with 0.63
England vs. Slovenia: England wins with 0.84
England vs. Serbia: England wins with 0.70
Denmark vs. Slovenia: Denmark wins with 0.67
Denmark vs. Serbia: Draw
Slovenia vs. Serbia: Serbia wins with 0.77
___Starting group E:___
Belgium vs. Romania: Belgium wins with 0.79
Belgium vs. Slovakia: Belgium wins with 0.79
Belgium vs. Ukraine: Belgium wins with 0.68
Romania vs. Slovakia: Slovakia wins with 0.60
Romania vs. Ukraine: Ukraine wins with 0.72
Slovakia vs. Ukraine: Ukraine wins with 0.71
___Starting group A:___
Germany vs. Hungary: Germany wins with 0.57
Germany vs. Scotland: Draw
Germany vs. Switzerland: Switzerland wins with 0.63
Hungary vs. Scotland: Scotland wins with 0.62
Hungary vs. Switzerland: Switzerland wins with 0.77
Scotland vs. Switzerland: Switzerland wins with 0.75
```

```
___Starting group F:___
Portugal vs. Turkey: Portugal wins with 0.69
Portugal vs. Czech Republic: Portugal wins with 0.66
Portugal vs. Georgia: Portugal wins with 0.85
Turkey vs. Czech Republic: Czech Republic wins with 0.63
Turkey vs. Georgia: Turkey wins with 0.63
Czech Republic vs. Georgia: Czech Republic wins with 0.66
___Starting group B:___
Spain vs. Albania: Spain wins with 0.80
Spain vs. Croatia: Croatia wins with 0.65
Spain vs. Italy: Italy wins with 0.63
Albania vs. Croatia: Croatia wins with 0.91
Albania vs. Italy: Italy wins with 0.91
Croatia vs. Italy: Italy wins with 0.59
___Starting group D:___
France vs. Austria: France wins with 0.75
France vs. Netherlands: Draw
France vs. Poland: France wins with 0.74
Austria vs. Netherlands: Netherlands wins with 0.81
Austria vs. Poland: Poland wins with 0.61
Netherlands vs. Poland: Netherlands wins with 0.63
```

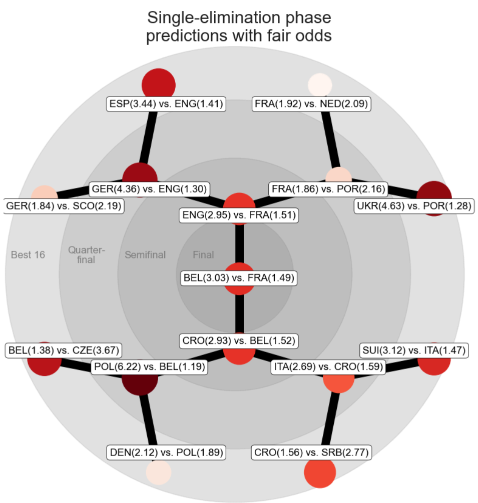
Calculate the points of each group based on the possibility predicted by the model. We will again rank them into the top 16 knockout rounds according to the real competition format (the top two in the group and the four best third places), and use the same model The results were predicted.

```
___Starting of the round_of_16___
Switzerland vs. Italy: Italy wins with probability 0.68
Croatia vs. Serbia: Croatia wins with probability 0.64
Denmark vs. Poland: Poland wins with probability 0.53
Belgium vs. Czech Republic: Belgium wins with probability 0.73
Germany vs. Scotland: Germany wins with probability 0.54
Spain vs. England: England wins with probability 0.71
France vs. Netherlands: France wins with probability 0.52
Ukraine vs. Portugal: Portugal wins with probability 0.78


___Starting of the quarterfinal___
Italy vs. Croatia: Croatia wins with probability 0.63
Poland vs. Belgium: Belgium wins with probability 0.84
Germany vs. England: England wins with probability 0.77
France vs. Portugal: France wins with probability 0.54


___Starting of the semifinal___
Croatia vs. Belgium: Belgium wins with probability 0.66
England vs. France: France wins with probability 0.66


___Starting of the final___
Belgium vs. France: France wins with probability 0.67
```



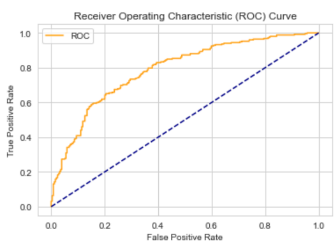
The training process and results of using different machine learning methods for the 2024 European Cup prediction model：

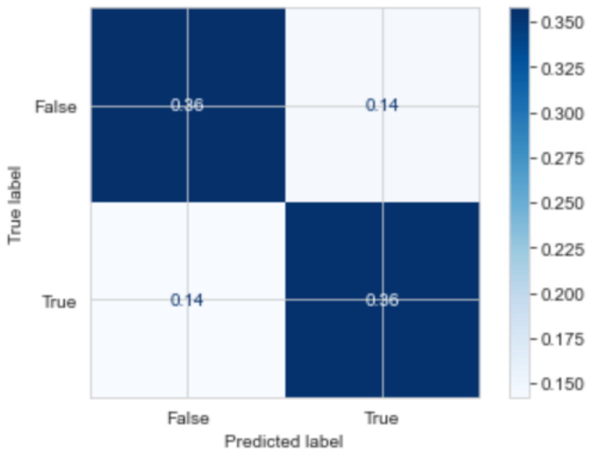
Logistic regression model training result:

```
Accuracy = 0.7134387351778656
ROC Area under Curve = 0.7134420697412824
Cohen's Kappa = 0.4268774703557312
Time taken = 0.022333145141601562
              precision    recall  f1-score   support

       False    0.71542   0.71260   0.71400       254
        True    0.71146   0.71429   0.71287       252

    accuracy                        0.71344       506
   macro avg    0.71344   0.71344   0.71344       506
weighted avg    0.71345   0.71344   0.71344       506
```



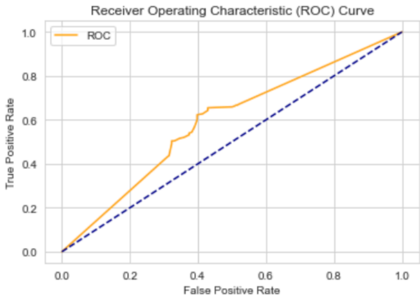


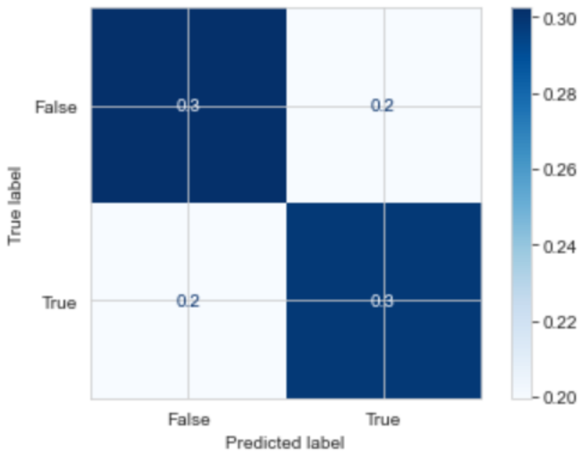
Decision tree model training result:

```
Accuracy = 0.6007905138339921
ROC Area under Curve = 0.6007842769653794
Cohen's Kappa = 0.20156855393075868
Time taken = 0.02096414566040039
              precision    recall  f1-score   support

       False    0.60236   0.60236   0.60236       254
        True    0.59921   0.59921   0.59921       252

    accuracy                        0.60079       506
   macro avg    0.60078   0.60078   0.60078       506
weighted avg    0.60079   0.60079   0.60079       506
```



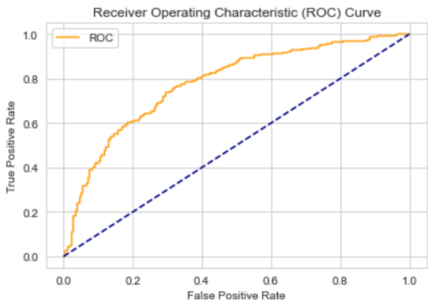


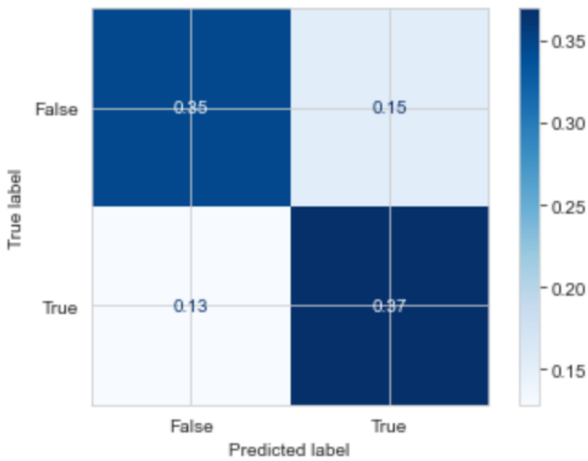
Neural network model training result:

```
Accuracy = 0.717391304347826
ROC Area under Curve = 0.7174884389451318
Cohen's Kappa = 0.4348885521938114
Time taken = 3.520681142807007
              precision    recall  f1-score   support

       False    0.73029   0.69291   0.71111       254
        True    0.70566   0.74206   0.72340       252

    accuracy                        0.71739       506
   macro avg    0.71798   0.71749   0.71726       506
weighted avg    0.71802   0.71739   0.71723       506
```



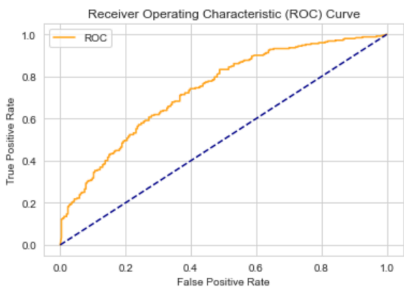


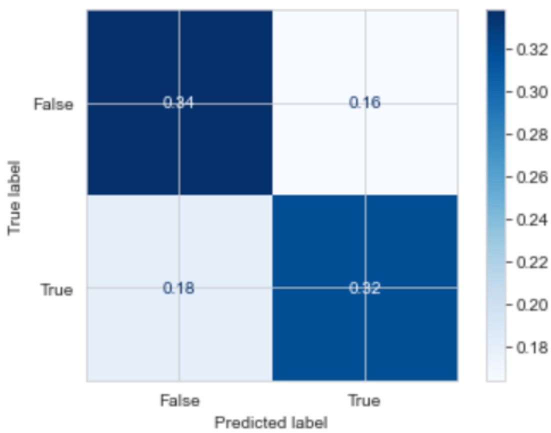
Random Forest model training result:

```
Accuracy = 0.6561264822134387
ROC Area under Curve = 0.656058617672791
Cohen's Kappa = 0.31215625
Time taken = 0.4522511959075928
              precision    recall  f1-score   support

       False    0.65267   0.67323   0.66279       254
        True    0.65984   0.63889   0.64919       252

    accuracy                        0.65613       506
   macro avg    0.65625   0.65606   0.65599       506
weighted avg    0.65624   0.65613   0.65602       506
```



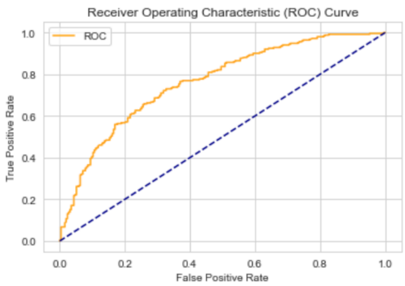


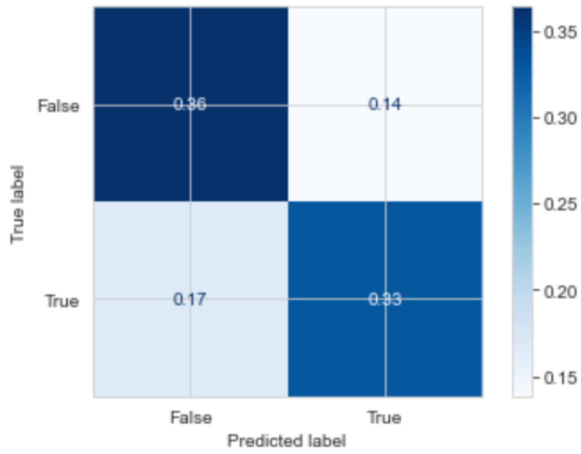
Lightgb model training result:

```
Accuracy = 0.6936758893280632
ROC Area under Curve = 0.6935539307586551
Cohen's Kappa = 0.3871985998468582
Time taken = 0.22465300559997559
              precision    recall  f1-score   support

       False    0.68401   0.72441   0.70363       254
        True    0.70464   0.66270   0.68303       252

    accuracy                        0.69368       506
   macro avg    0.69433   0.69355   0.69333       506
weighted avg    0.69429   0.69368   0.69337       506
```



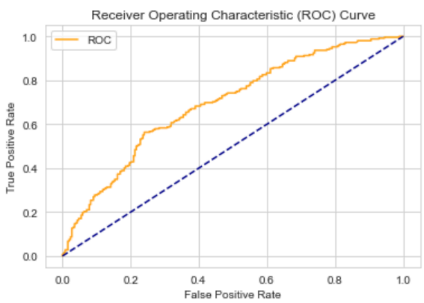


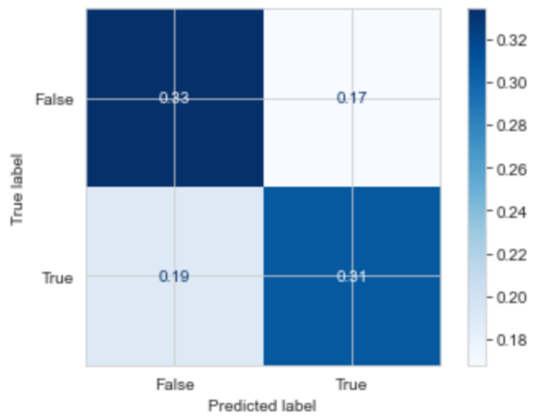
Xgboost model training result:

```
Accuracy = 0.642292490118577
ROC Area under Curve = 0.6422009748781402
Cohen's Kappa = 0.28445083363282664
Time taken = 3.3314640522003174
              precision    recall  f1-score   support

       False    0.63774   0.66535   0.65125       254
        True    0.64730   0.61905   0.63286       252

    accuracy                        0.64229       506
   macro avg    0.64252   0.64220   0.64206       506
weighted avg    0.64250   0.64229   0.64209       506
```

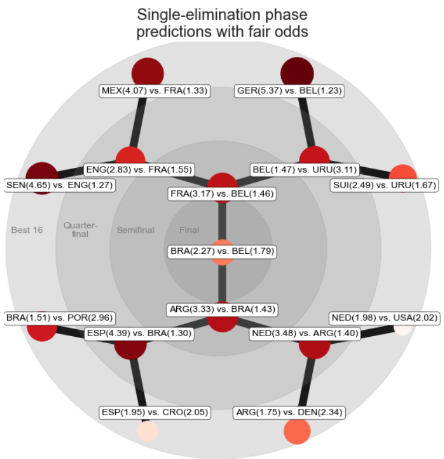




F. Discussion

The discussion of the prediction results is divided into two parts. First, we analyze the match outcomes of the 2022 World Cup to validate our model; second, we examine the predicted results for the 2024 European Championship of different models.

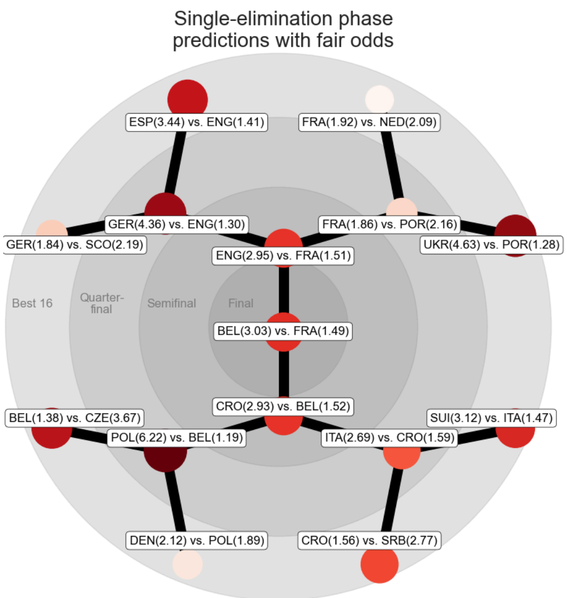
For the analysis of the predicted outcomes of the 2022 World Cup:



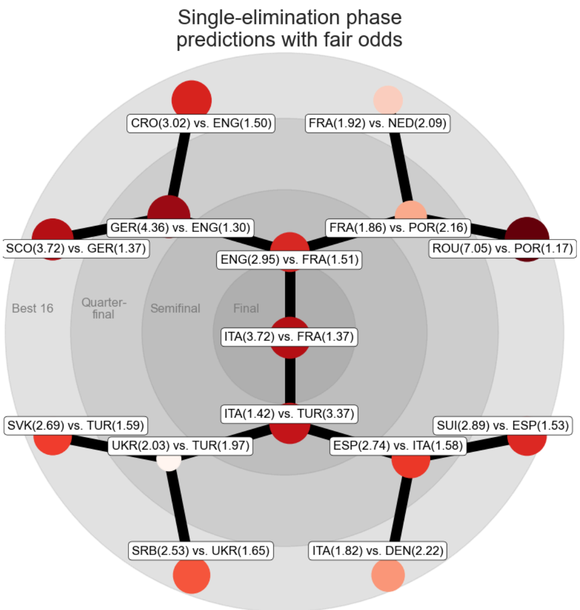
Our model correctly predicted multiple outcomes, such as: in the round of 16, our model successfully predicted 11 teams, including the Netherlands, Argentina, the United States, among others. We also accurately forecasted some match results, such as Argentina winning against New Zealand, and England winning against Senegal. Therefore, I believe our model has a certain degree of reliability.

For the analysis of the predicted outcomes of the 2024 European Championship:

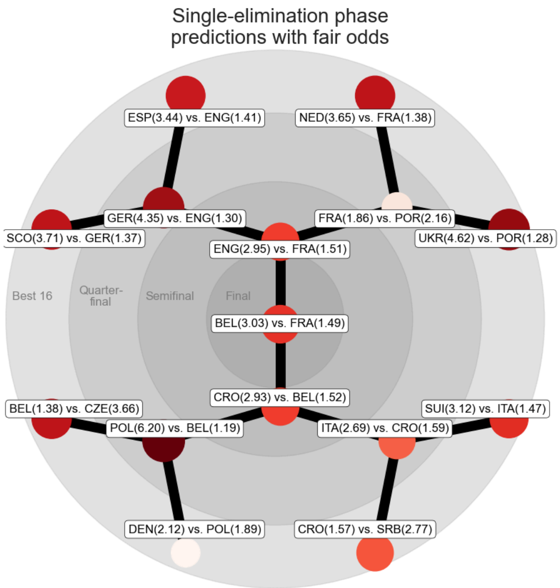
Our model predicts the round of 16 teams to be: Switzerland, Italy, Croatia, Serbia, Denmark, Poland, Belgium, Czech Republic, Germany, Scotland, Spain, England, France, Netherlands, Ukraine, Portugal. The quarter-finalists are: Italy, Croatia, Belgium, Poland, Germany, England, France, Portugal. The semi-finalists are: Croatia, Belgium, England, France. The final champion is predicted to be France.



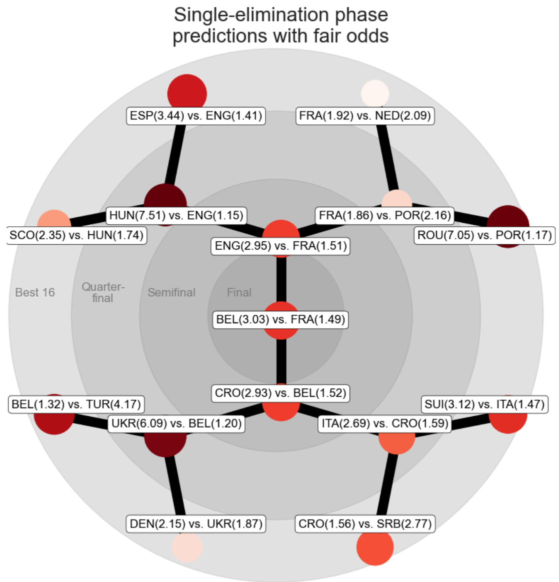
Logistic regression prediction results



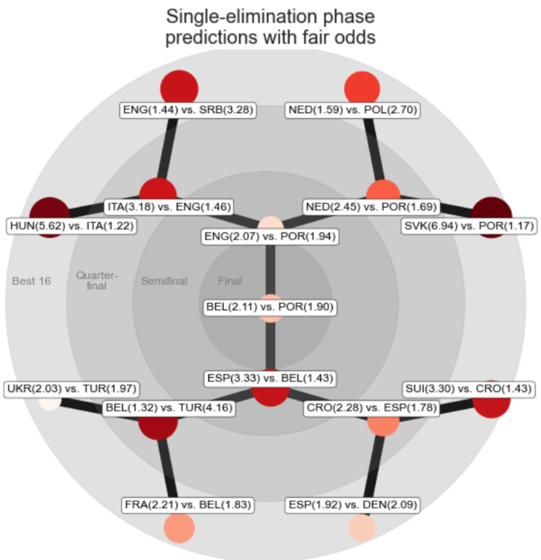
Decision tree prediction results



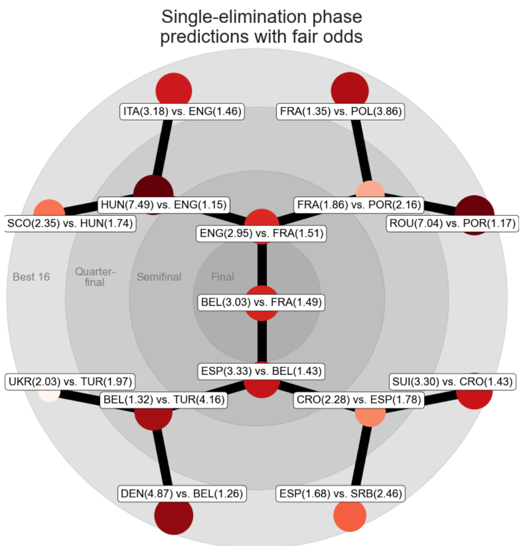
Neural network prediction results



Random forest prediction results



Lightgbm prediction results



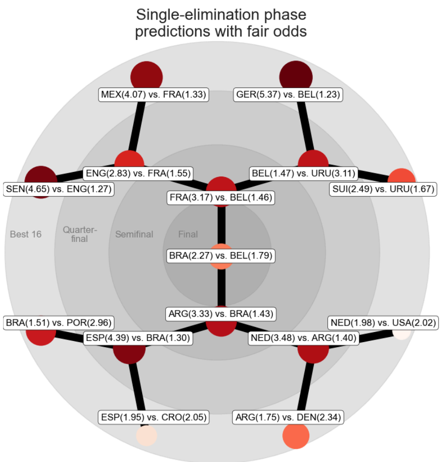
Xgboost prediction results

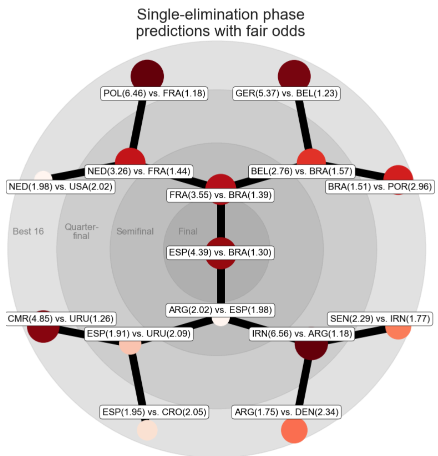
2024 European Cup prediction results

	Logistic Regression	Decision Tree	MLP	Random Forest	LGB	XGB
Champion	France	France	France	France	Portugal	France
Runner-up	Belgium	Italy	Belgium	Belgium	Belgium	Belgium

2022 World Cup prediction results and model verification

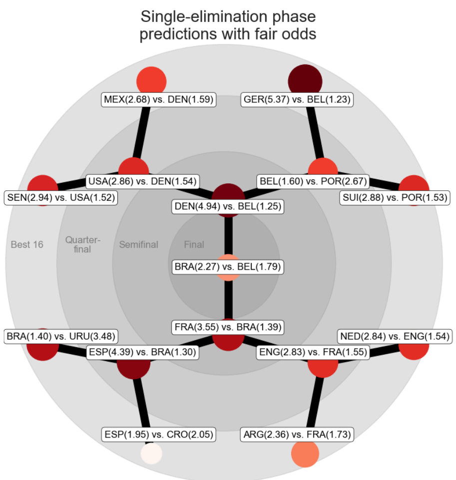
	Logistic Regression	Decision Tree	MLP	Random Forest	LGBM	XGB	Combinition
Prediction accuracy rate of final Four	50%	50%	25%	50%	50%	50%	50%
Prediction accuracy rate of the top eight	62.5%	50%	50%	62.5%	62.5%	62.5%	62.5%
Prediction accuracy rate of top 16	68.75%	62.5%	68.75%	56.25%	62.5%	56.25%	75%

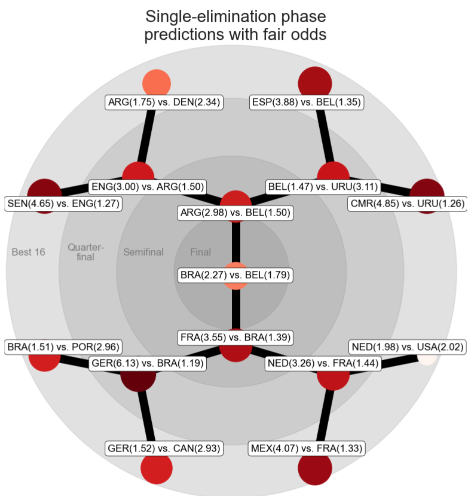




Logistic regression prediction results

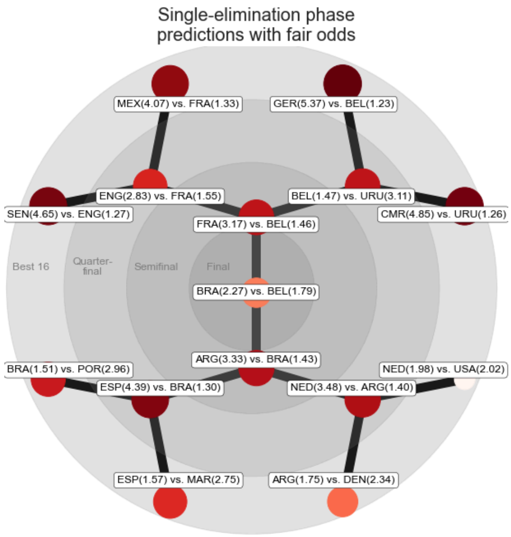
Decision tree prediction results



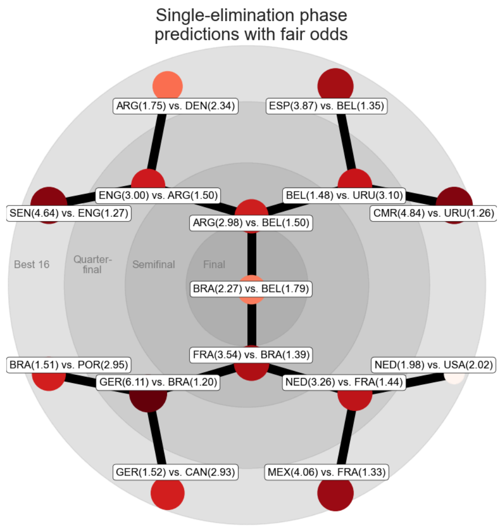


Neural network prediction results

Random forest prediction results



Lightgbm prediction results



Xgboost prediction results

G. Future Work

Although this work has predicted the outcomes of the 2024 European Championship, its validity and accuracy cannot be verified as of yet, even though we have already validated the model with the results of the 2022 World Cup. Additionally, due to dataset limitations, the features we can use are limited, primarily focusing on the historical performance of teams. In the future, as data becomes more abundant, I believe future work may take into consideration more relevant data, such as player conditions: the health and status of key players, and whether any are injured or suspended. Weather conditions: for instance, rain might affect ball control and pitch conditions, thereby influencing match outcomes. Schedule density and fatigue: the team's scheduling, and the density of recent matches could lead to player fatigue, affecting their performance, and so on.

H. Conclusion

By delving deeply into international football match data from the past several decades and the world rankings of national teams, and training different models on this data, we can develop predictive models for forecasting future match outcomes. To ensure the reliability of our training data, we preprocessed the data to address gaps and conducted correlation checks. We also adjusted the data dimensions to reduce the likelihood of overfitting, gradient exploding, or gradient vanishing in our models to some extent.

We validated the completed models on two major events (the already occurred 2022 World Cup and the upcoming 2024 European Championship). We found that, in the predictions for the 2022 World Cup, despite many dark horses and surprising defeats of strong teams, we still demonstrated relatively good predictive results for the round of 16 and quarter-finals.

Given the immense unpredictability of World Cup matches, our model can be verified to have unearthed the inherent relationships and correlations within the data.

Since we trained six different models and all made predictions on match outcomes, integrating the results of the six models (where the six models vote, and the majority's choice is taken as the prediction result) showed to have better predictive accuracy.

I. References

[1] Jawade I, Jadhav R, Vaz M, et al. Predicting Football Match Results using Machine Learning[J]. Maharashtra: International Research Journal of Engineering and Technology, 2021.
[2] Penn M J, Donnelly C A. Analysis of a double Poisson model for predicting football results in Euro 2020[J]. Plos one, 2022, 17(5): e0268511.
[3] Ren Y, Susnjak T. Predicting football match outcomes with explainable machine learning and the kelly index[J]. arXiv preprint arXiv:2211.15734, 2022.
[4] Carloni, Luca & De Angelis, Andrea & Sansonetti, Giuseppe & Micarelli, Alessandro. (2021). A Machine Learning Approach to Football Match Result Prediction. 10.1007/978-3-030-78642-7_63.

J. Annex

Topic Choice: Shuhao Liu, Heshun Wang, Chenyu Song, Jianpeng Chen
Data Collecting: Heshun Wang, Chenyu Song
Data processing(Data Cleaning, Dimensionality Reduction, etc.): Shuhao Liu, Jianpeng Chen, Heshun Wang
World Cup and European Cup Promotion rules Algorithm: Heshun Wang
Model Training and evaluation: Chenyu Song
Results analysis: Jianpeng Chen
Report: Chenyu Song, Jianpeng Chen
Presentation: Shuhao Liu
PPT: Shuhao Liu, Heshun Wang