ICCS261 Term Project: Predicting

—

Kan-Anek Tantitayapong (6380612)

Data Collection

Datasets used

Dataset 1: International Football Results 1872-2024 (results.csv and shootouts.csv)
https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017

Dataset 2: FIFA World Ranking 1992-2024 (rankings.csv)
https://www.kaggle.com/datasets/cashncarry/fifaworldranking

results.csv

```
results = pd.read_csv("results.csv")
results.head()
```

✓ 0.0s

	date	home_team	away_team	home_score	away_score	tournament	city	country	neutral
0	1872-11-30	Scotland	England	0.0	0.0	Friendly	Glasgow	Scotland	False
1	1873-03-08	England	Scotland	4.0	2.0	Friendly	London	England	False
2	1874-03-07	Scotland	England	2.0	1.0	Friendly	Glasgow	Scotland	False
3	1875-03-06	England	Scotland	2.0	2.0	Friendly	London	England	False
4	1876-03-04	Scotland	England	3.0	0.0	Friendly	Glasgow	Scotland	False

shootouts.csv

```
shootouts = pd.read_csv("shootouts.csv")
shootouts.tail()
```

✓ 0.0s

	date	home_team	away_team	winner	first_shooter
631	2024-03-26	New Zealand	Tunisia	Tunisia	New Zealand
632	2024-03-26	Wales	Poland	Poland	Poland
633	2024-03-26	Georgia	Greece	Georgia	Georgia
634	2024-03-26	Turks and Caicos Islands	Anguilla	Anguilla	Turks and Caicos Islands
635	2024-03-26	British Virgin Islands	United States Virgin Islands	British Virgin Islands	British Virgin Islands

rankings.csv

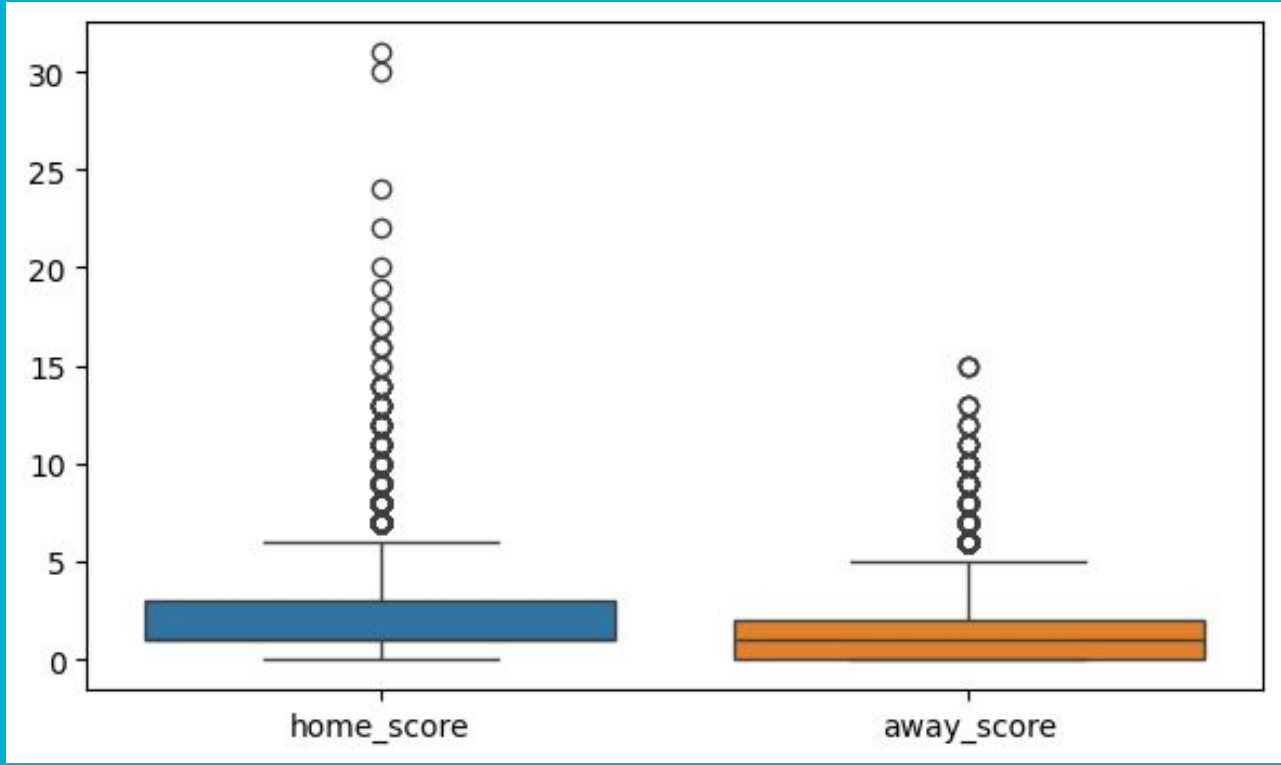
```
rankings = pd.read_csv("rankings.csv")
rankings.head()
```

✓ 0.0s

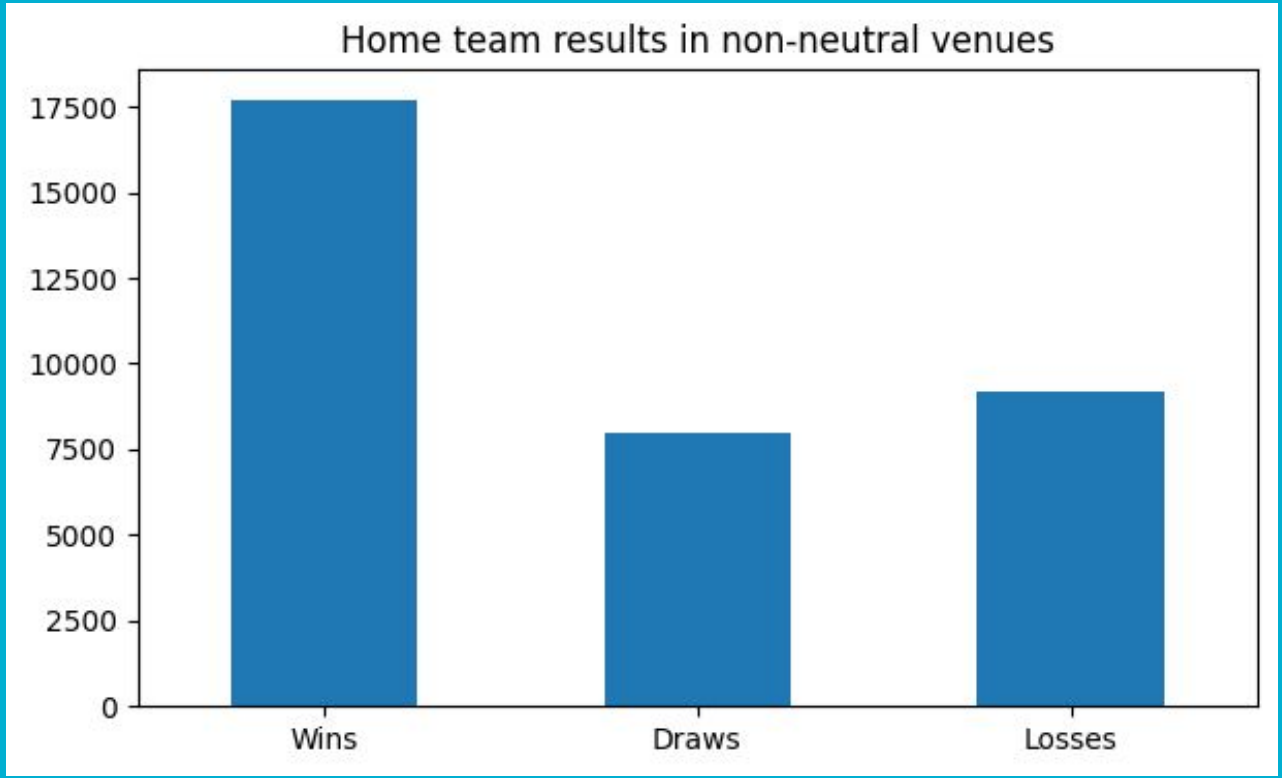
	rank	country_full	country_abrv	total_points	previous_points	rank_change	confederation	rank_date
0	83.0	Guatemala	GUA	15.0	0.0	83	CONCACAF	1992-12-31
1	32.0	Zambia	ZAM	38.0	0.0	32	CAF	1992-12-31
2	33.0	Portugal	POR	38.0	0.0	33	UEFA	1992-12-31
3	34.0	Austria	AUT	38.0	0.0	34	UEFA	1992-12-31
4	35.0	Colombia	COL	36.0	0.0	35	CONMEBOL	1992-12-31

Exploratory Data Analysis

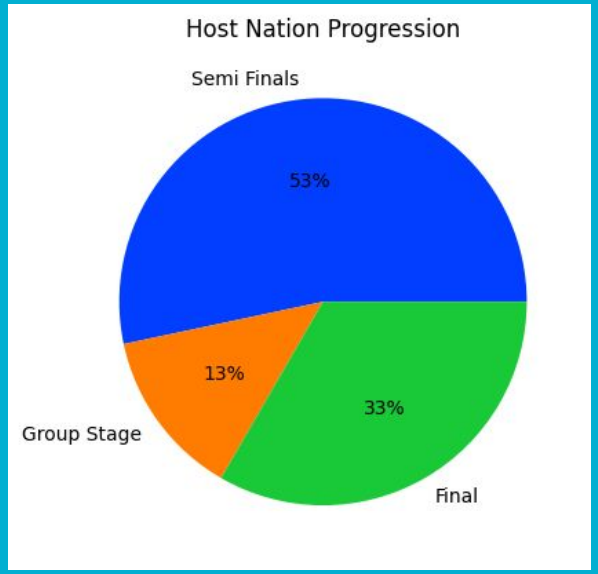
Goals scored by home teams vs away teams

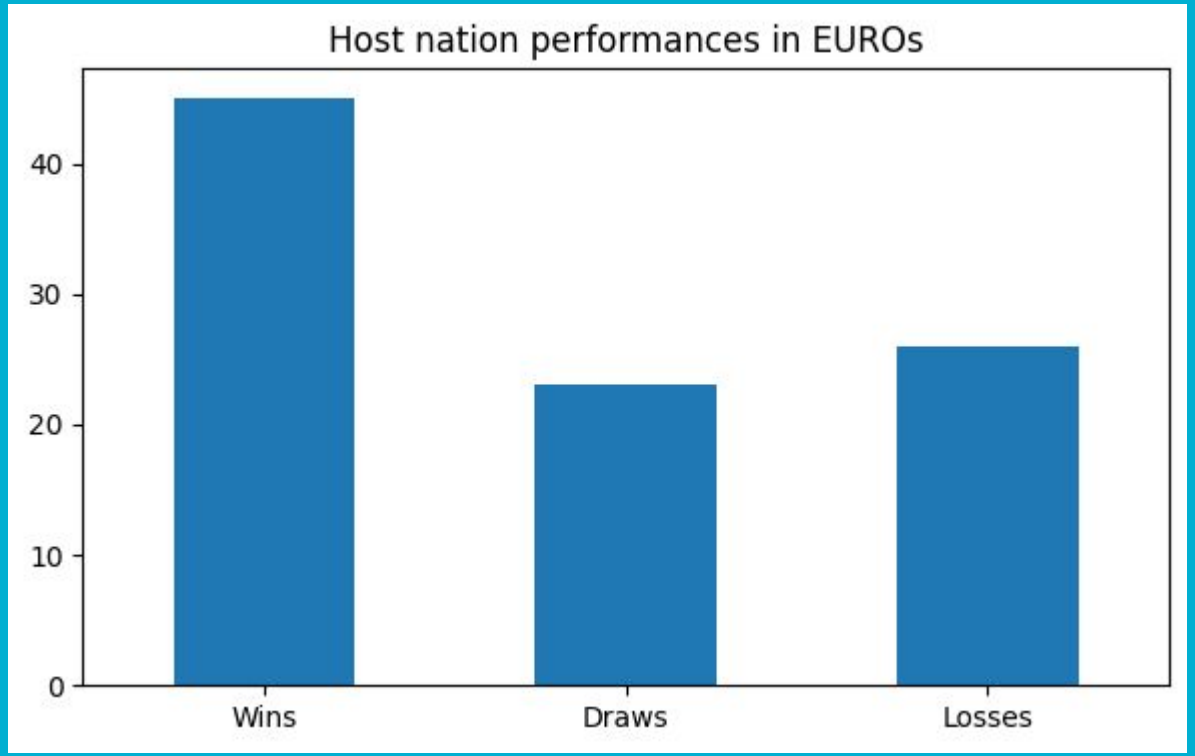


How teams perform when playing at home

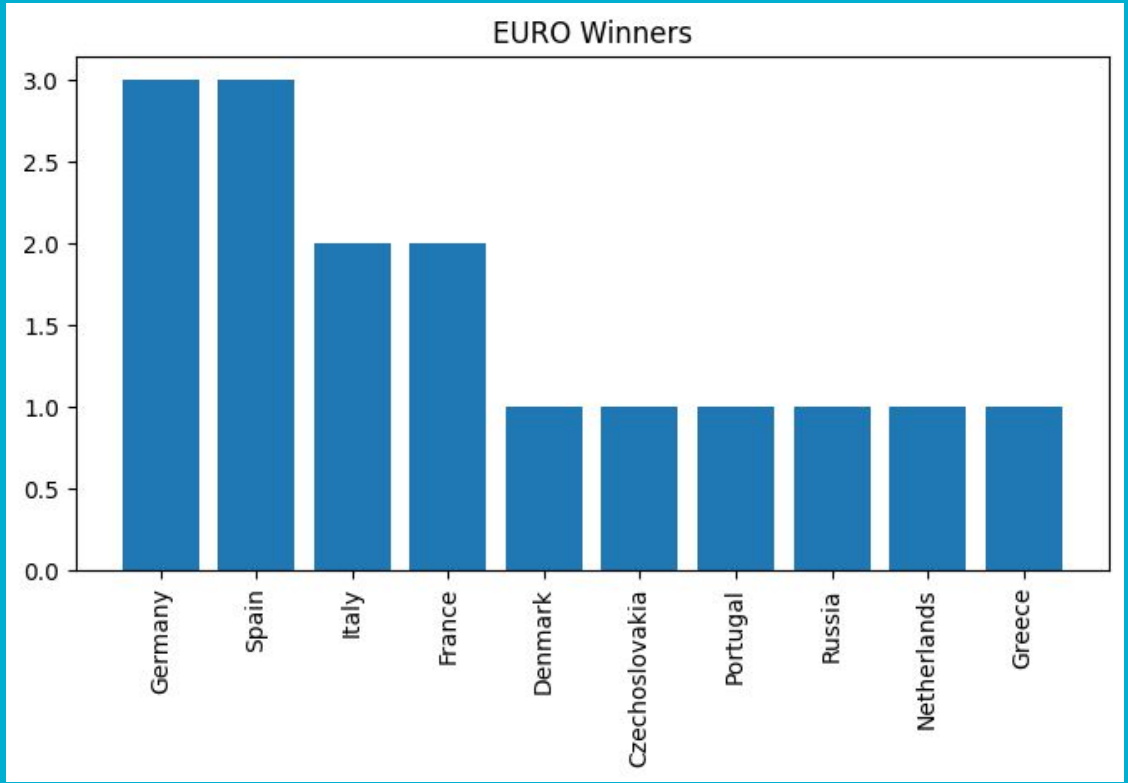


How host nations perform at the Euros

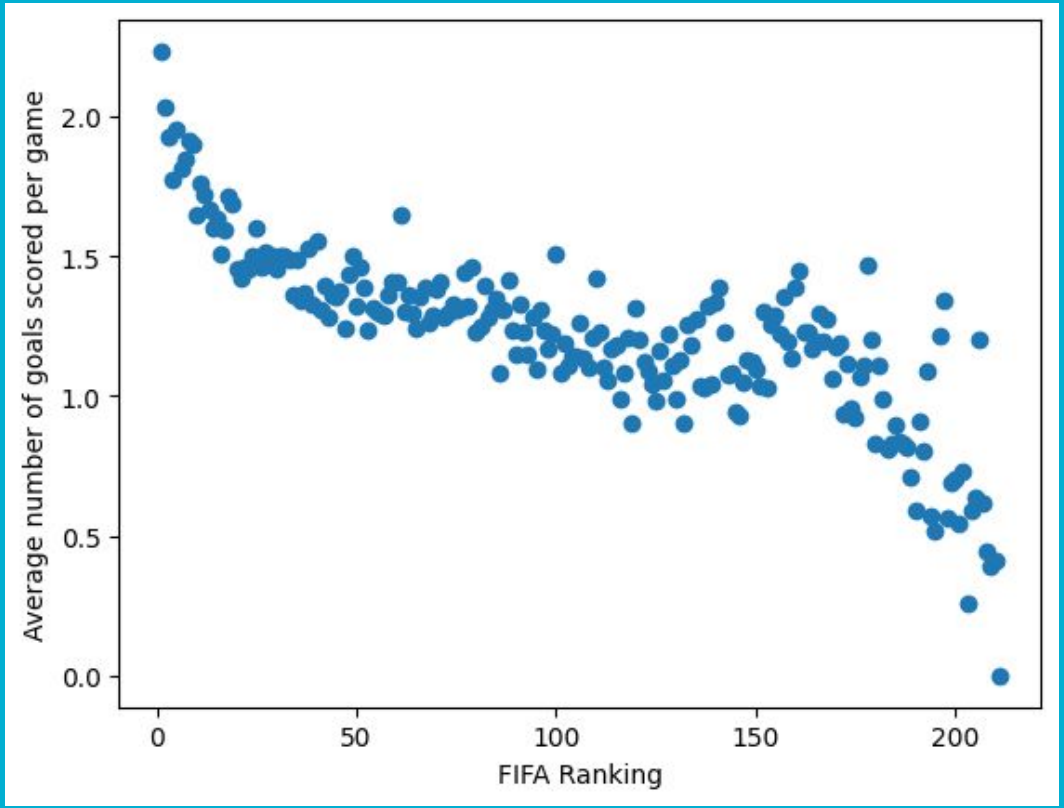




Past winners



FIFA ranking vs average goals scored



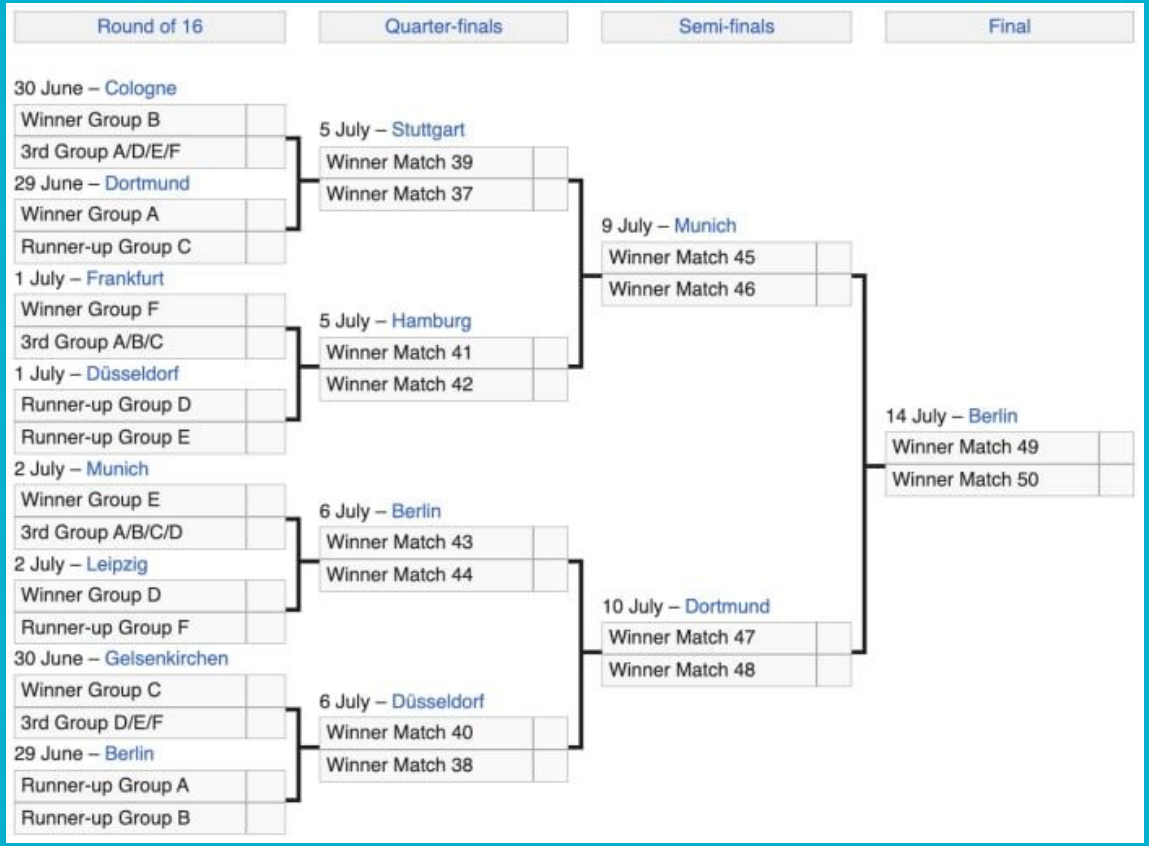
r^2=0.6834

Model Development

Preliminaries

- The Euros is a tournament-style competition. The first stage is the group stage, where the 24 teams are divided into 6 groups of 4 teams each. Every team in the group plays each other exactly once and accumulates points.
- Win: 3pts, Draw: 1pt, Lose: 0pts
- The top two teams of each group qualify for the round of 16.
- Furthermore, the four best third place teams also qualify for the round of 16.
- If two teams have the same number of points, the tiebreaker is first the head to head, then goal difference, then number of goals scored, then fair play record.

Bracket


Round of 16
Quarter-finals
Semi-finals
Final
30 June – Cologne
Winner Group B
3rd Group A/D/E/F
29 June – Dortmund
Winner Group A
Runner-up Group C
1 July – Frankfurt
Winner Group F
3rd Group A/B/C
1 July – Düsseldorf
Runner-up Group D
Runner-up Group E
2 July – Munich
Winner Group E
3rd Group A/B/C/D
2 July – Leipzig
Winner Group D
Runner-up Group F
30 June – Gelsenkirchen
Winner Group C
3rd Group D/E/F
29 June – Berlin
Runner-up Group A
Runner-up Group B
5 July – Stuttgart
Winner Match 39
Winner Match 37
5 July – Hamburg
Winner Match 41
Winner Match 42
6 July – Berlin
Winner Match 43
Winner Match 44
6 July – Düsseldorf
Winner Match 40
Winner Match 38
9 July – Munich
Winner Match 45
Winner Match 46
10 July – Dortmund
Winner Match 47
Winner Match 48
14 July – Berlin
Winner Match 49
Winner Match 50

Features

- Goal difference is significant to a team's position in a group, and hence their position in the bracket. So, for a match where Team A plays Team B, we want to predict both Team A's score and Team B's score.
- Both teams' FIFA ranking will be used as two features, as it is a key determinant to a team's strength.
- Whether or not the venue they play in is neutral (0 or 1) will also be a feature, since playing in front of a home crowd serves to give the team a boost.

Supporting Model

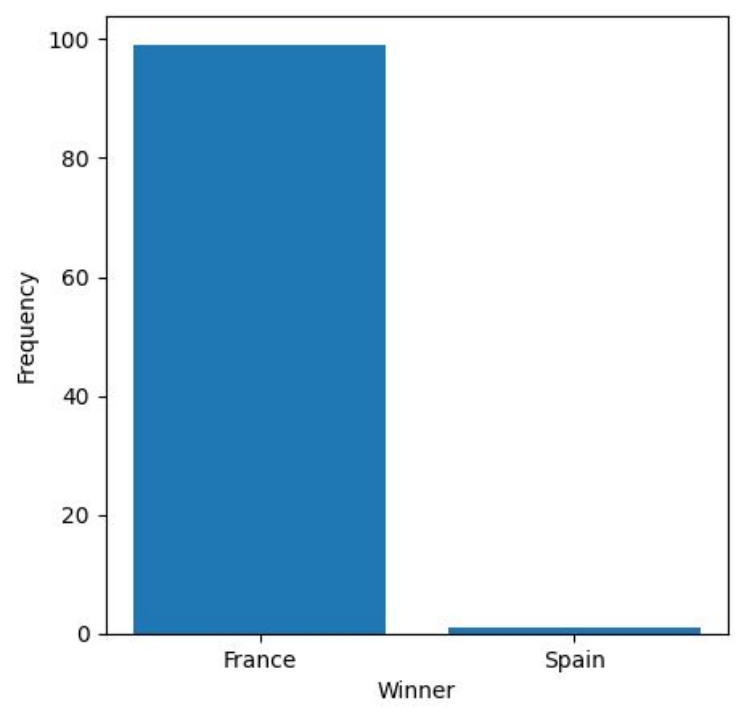
- When a knockout game ends in a tie, there will be 30 more minutes added on. If after the 30 minutes the game is still tied, then the winner is decided using a penalty shootout. So, whenever our main model predicts a tie in the knockout phase, we will use this model to break the tie.
- The model is a Logistic Regression model which given the FIFA rankings of the teams and whether or not the venue is neutral, predicts the winner of shootout. The model is trained on the shootouts.csv dataset.
- It must be noted that penalty shootouts are very hard to predict quantitatively, because it mostly comes down to which team can hold their nerves better.

Model 1: Linear Regression

After 100 trials:

France: 99

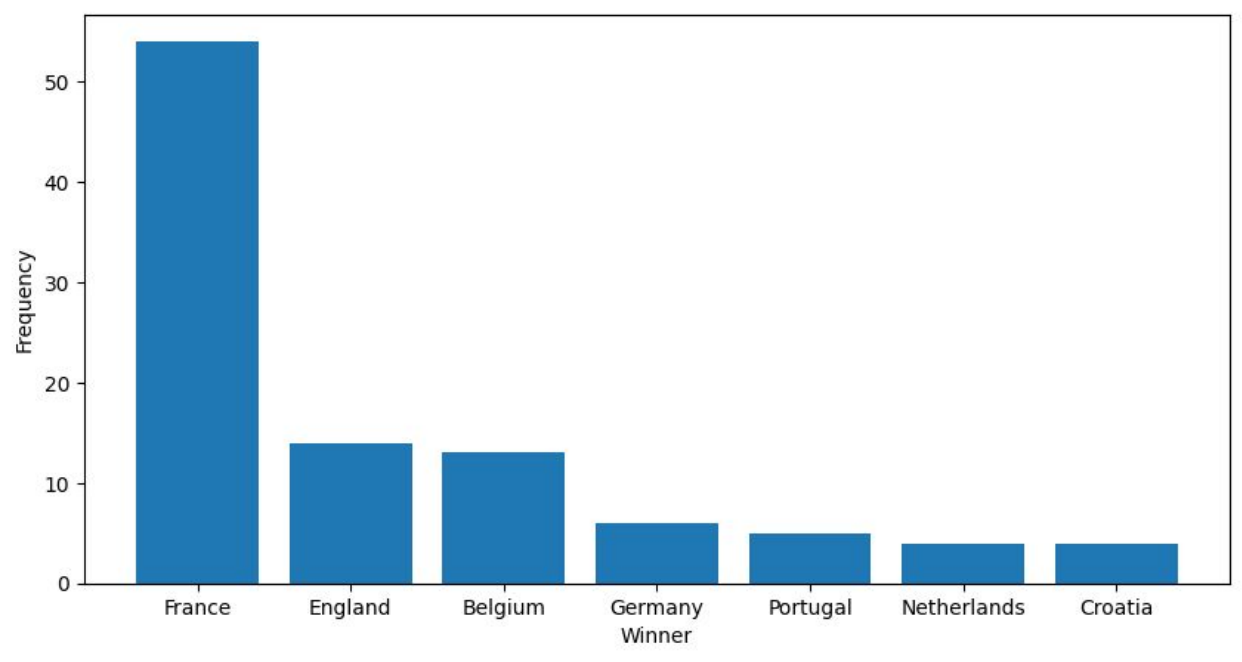
Spain: 1



Model is very biased towards the FIFA ranking of the team, since France (2) is the highest ranked team in the competition

Model 2: Random Forest

More variability than Model 1. We see that there is still bias towards the FIFA ranking but also some significance is given to Germany (ranked 16) hosting this year's competition.

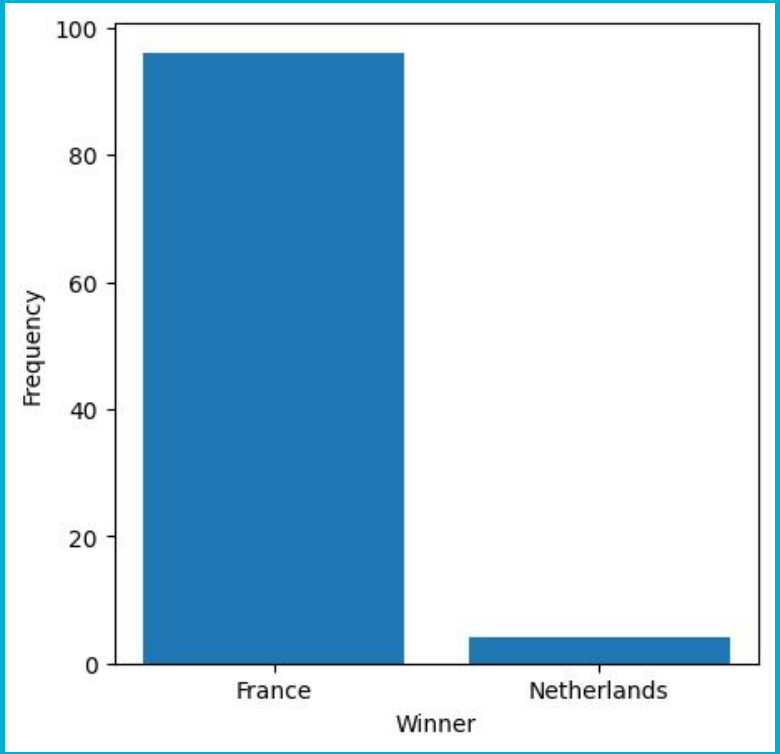


Model 3: Neural Networks

After 100 trials:

France: 96

Netherlands: 4



Game-by-Game Simulation With Random Forest Model

Hyperparameters

```
param_grid = {
    'estimator__n_estimators': [100, 200, 300],
    'estimator__max_depth': [None, 10, 20],
    'estimator__min_samples_split': [2, 5, 10],
    'estimator__min_samples_leaf': [1, 2, 4],
    'estimator__max_features': ['auto', 'sqrt', 'log2']
}
```

After a quick grid search, we find that the higher number of estimators, the lower the error. Thus, we will be using 500 estimators with 70% train and 30% test. Also, fixing more hyperparameters makes the model increasingly deterministic. Because of this we will only be fixing the number of estimators.

Group Stages

Group A

		Model	Actual	
GER	VS SCO	3–2	5–1	✅
HUN	VS SUI	1–2	1–3	✅
GER	VS HUN	2–1	2–0	✅
SCO	VS SUI	1–1	1–1	✅ ✅
GER	VS SUI	1–1	1–1	✅ ✅
SCO	VS HUN	0–1	0–1	✅ ✅

Group Table Comparison

Model	Team	Actual	Difference
1	GER	1	0
2	SUI	2	0
3	HUN	3	0
4	SCO	4	0

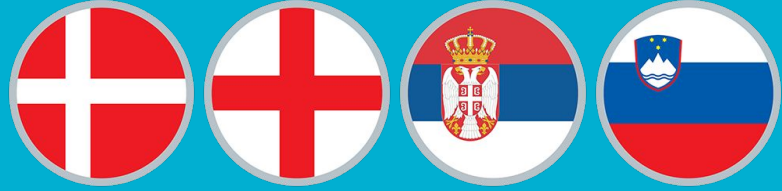
Group B

	Model	Actual	
VS	1–1	3–0	❌
VS	2–1	2–1	✅✅
VS	2–1	2–2	❌
VS	1–0	1–0	✅✅
VS	1–2	0–1	✅
VS	1–2	1–1	❌

Group Table Comparison

Model	Team	Actual	Difference
1	ESP	1	0
2	ITA	2	0
3	CRO	3	0
4	ALB	4	0

Group C

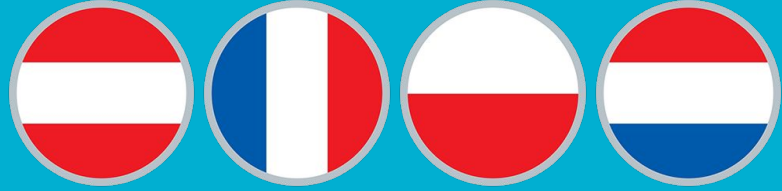


		Model	Actual	
SVN	VS DEN	1–1	1–1	✅✅
SRB	VS ENG	0–2	0–1	✅
SVN	VS SRB	1–2	1–1	❌
DEN	VS ENG	1–1	1–1	✅✅
DEN	VS SRB	2–1	0–0	❌
ENG	VS SVN	1–0	0–0	❌

Group Table Comparison

Model	Team	Actual	Difference
1	ENG	1	0
2	DEN	2	0
4	SVN	3	+1
3	SRB	4	-1

Group D



	Model	Actual	
POL VS NED	1–2	1–2	✅✅
AUT VS FRA	2–1	0–1	❌❌
POL VS AUT	1–2	1–3	✅
NED VS FRA	1–1	0–0	✅
NED VS AUT	2–1	2–3	❌❌
FRA VS POL	2–1	1–1	❌

Group Table Comparison

Model	Team	Actual	Difference
2	AUT	1	+1
3	FRA	2	+1
1	NED	3	-2
4	POL	4	0

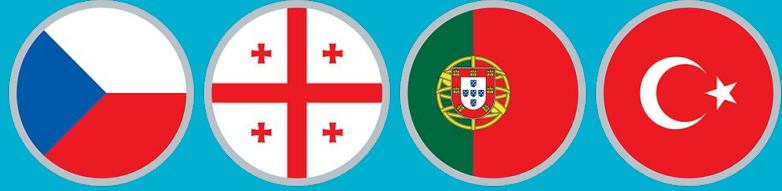
Group E

		Model	Actual	
BEL	VS SVK	1–2	0–1	✅
ROU	VS UKR	1–1	3–0	❌
SVK	VS UKR	2–0	1–2	❌❌
BEL	VS ROU	2–1	2–0	✅
SVK	VS ROU	0–1	1–1	❌
UKR	VS BEL	1–2	0–0	❌

Group Table Comparison

Model	Team	Actual	Difference
3	ROU	1	+2
2	BEL	2	0
1	SVK	3	-2
4	UKR	4	0

Group F



		Model	Actual	
TUR	VS GEO	2–2	3–1	❌
POR	VS CZE	1–1	2–1	❌
GEO	VS CZE	1–1	1–1	✅✅
TUR	VS POR	0–2	0–3	✅
GEO	VS POR	1–3	2–0	❌❌
CZE	VS TUR	2–1	1–2	❌❌

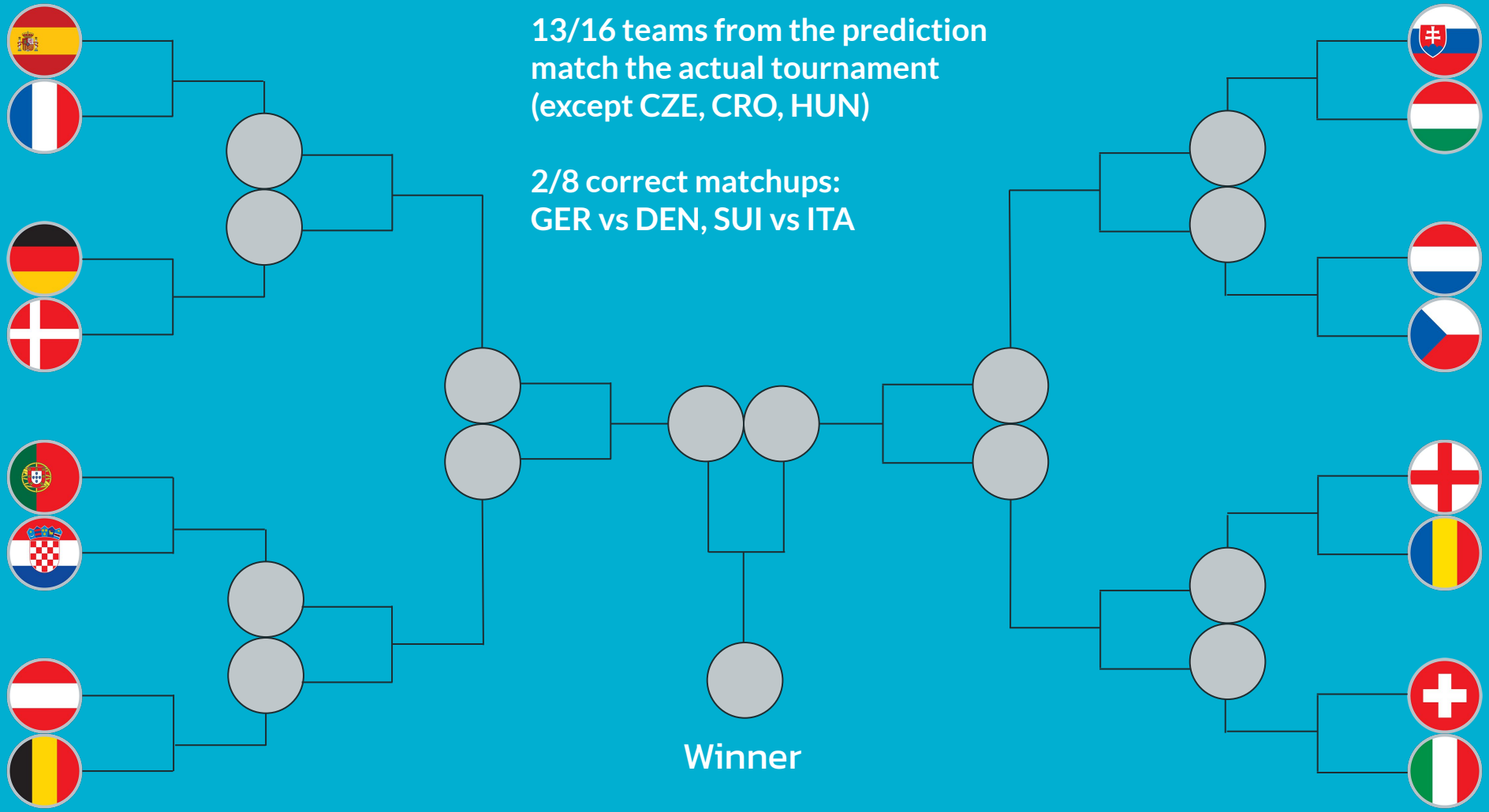
Group Table Comparison

Model	Team	Actual	Difference
1	POR	1	0
4	TUR	2	+2
3	GEO	3	0
2	CZE	4	-2

Knockouts

13/16 teams from the prediction match the actual tournament (except CZE, CRO, HUN)

2/8 correct matchups:
GER vs DEN, SUI vs ITA

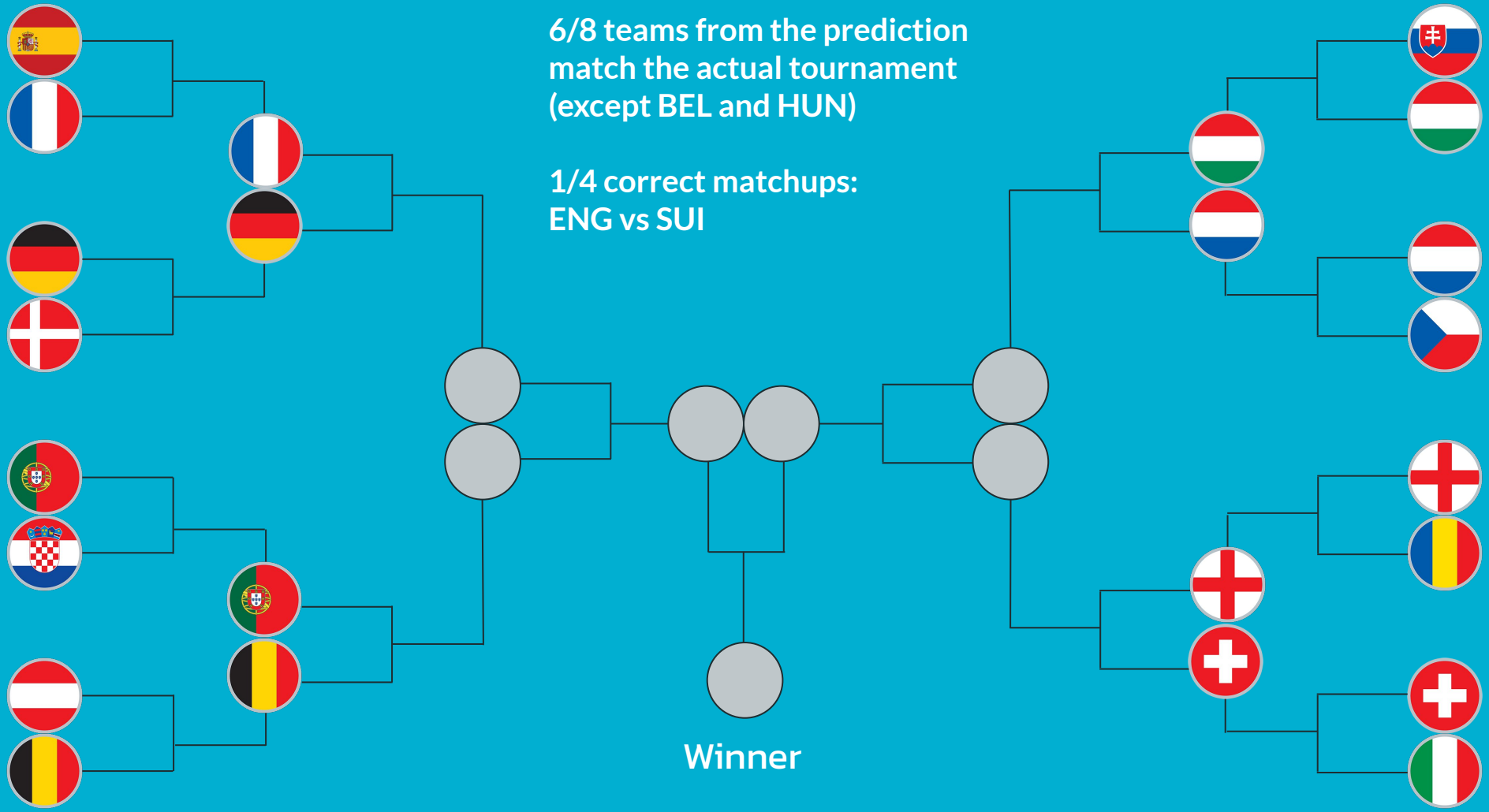


Round of 16 Prediction

	Prediction			Prediction
VS	2–2, FRA on pens		VS	1–1, HUN on pens
VS	2–0	✅✅	VS	2–1
VS	1–1, POR on pens		VS	2–1
VS	1–2		VS	2–2, SUI on pens ✅

6/8 teams from the prediction match the actual tournament (except BEL and HUN)

1/4 correct matchups:
ENG vs SUI

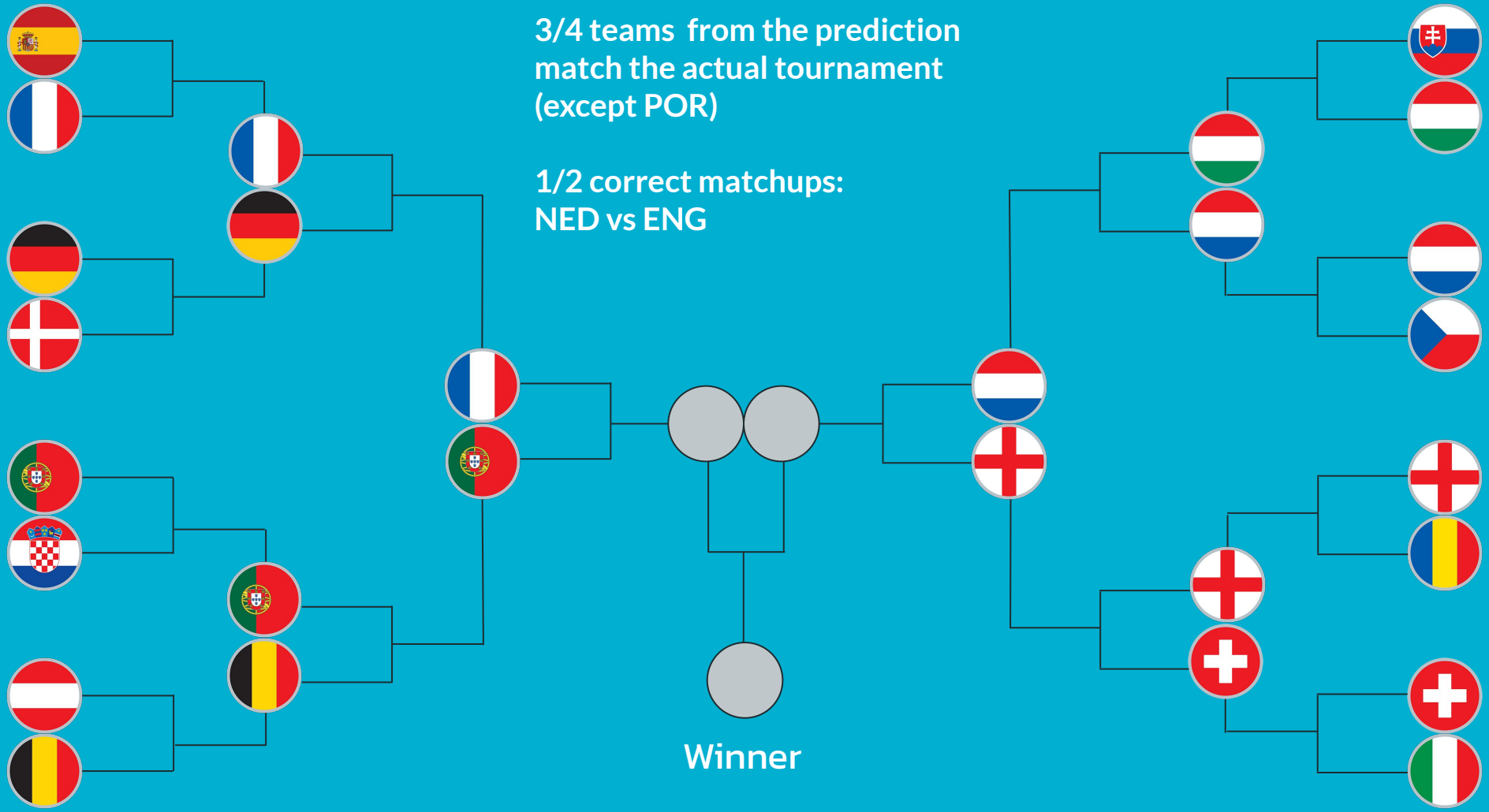


Quarter-finals Prediction

	Prediction
VS	1–0
VS	1–1, POR on pens
VS	1–2
VS	2–0 ✅

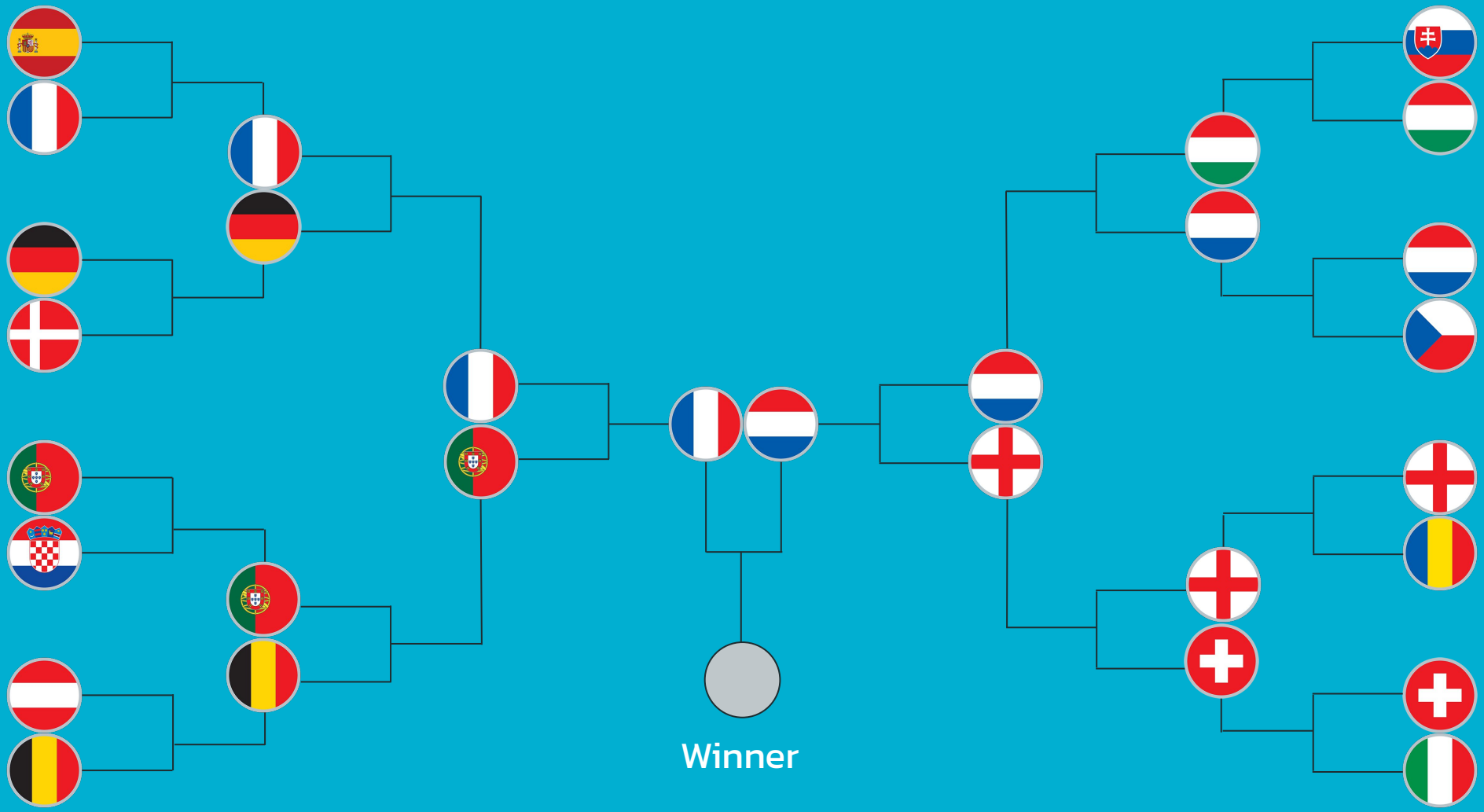
3/4 teams from the prediction match the actual tournament (except POR)

1/2 correct matchups:
NED vs ENG



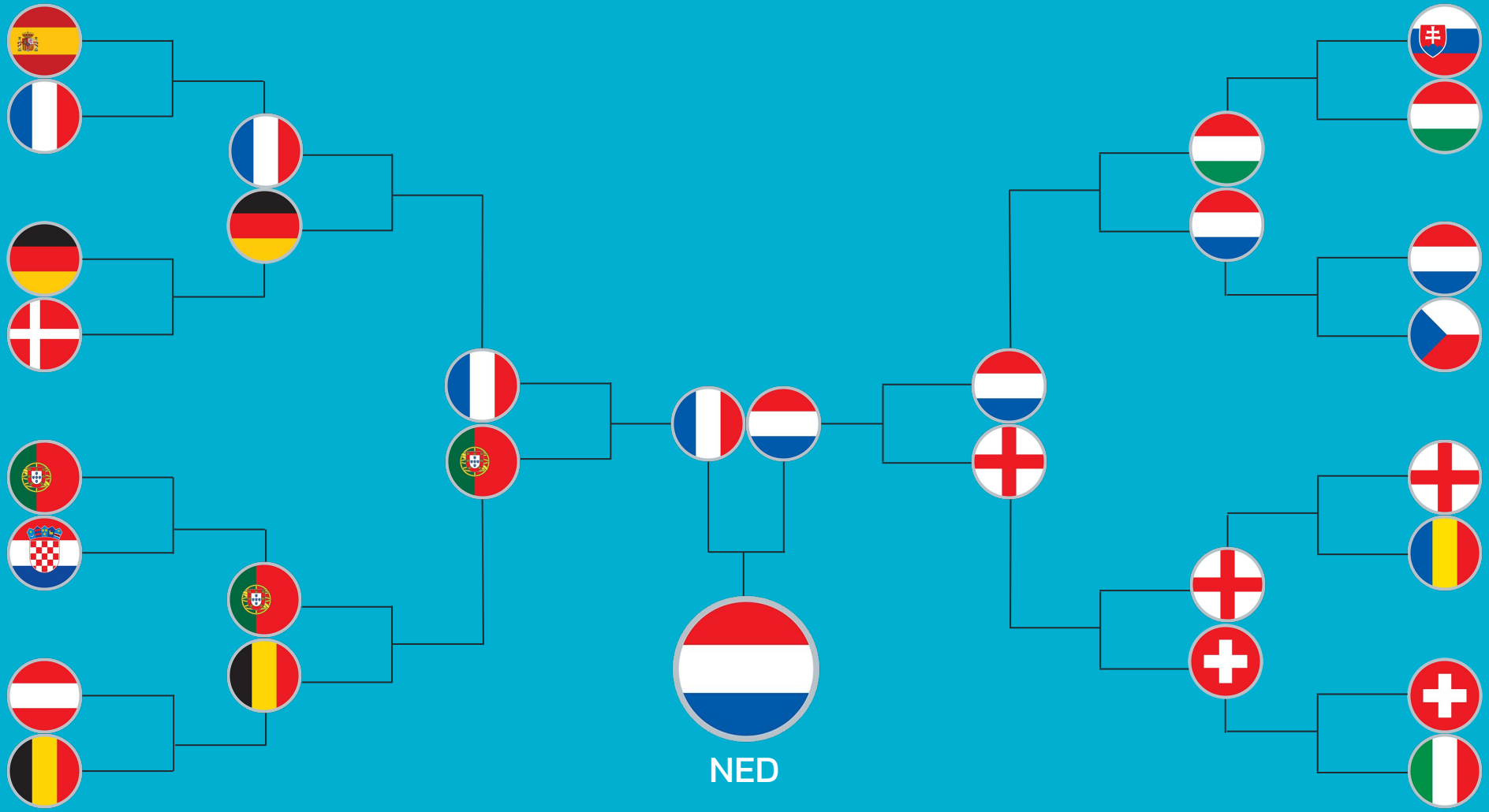
Semi-finals Prediction

			Prediction
	VS		1–1, FRA on pens
	VS		2–1


Winner

Final Prediction

1 2


NED