University of Halabja

College of Science

Computer Science Department

Data Mining Coursework Description

FIFA World Cup 2022

Group Stage Prediction

Prepared by:

Dadyar hamaxald Karzan kamaran

Superviser :

Mohammed L. Mahmood

Table of Contents

FIFA World Cup 2022 Group Stage Prediction

The FIFA World Cup is the most prestigious football tournament in the world. The championship has been awarded every four years since the start of the tournament in 1930.

The current format involves a qualification phase, which takes place over the preceding three years, to determine which teams quality for the tournament. In the tournament, 32 teams, including the host nation, compete for the title at different stadiums in the host country.

The reigning champion is France, which beat Croatia in the 2018 tournament in Russia. Qatar will host the 2022 tournament, for which the first match will be played in November.

This dataset provides a complete overview of all international soccer matches played since the 90s. On top of that, the strength of each team is provided by incorporating actual FIFA rankings as well as player strengths and team.

Some suggestions

· Can you predict what team is most likely to win the 2022 FIFA World Cup?

· What team has the strongest defense, midfield, and offense players?

· Is there really such a thing as a home team advantage?

· Do teams with stronger offense players score more goals? And do teams with stronger goalkeepers receive fewer goals?

· What team has the longest winning streak?

· Does the best team always win? Can you explain why a weaker team sometimes win?

Online Dataset collection

Due to the lack of fresh data, we have tried to use old data for predictions.

referenced some awesome Kaggle notebooks for creating this, here are the references.

- ✓ https://www.kaggle.com/code/agostontorok/soccer-world-cup-2018-winner/notebook
- ✓ https://www.kaggle.com/code/startupsci/titanic-data-science-solutions

Data Pre-processing

Feature Selection

Here we have explained the data available to us that cause problems or ambiguity in the results.

Example:

```
Replace >>> {"IR Iran": "Iran", "Korea Republic" : "South Korea"}
```

```
match_df = match_df.replace({"IR Iran": "Iran", "Korea Republic" : "South Korea"})
rank_df = rank_df.replace({"IR Iran": "Iran", "Korea Republic" : "South Korea"})
```

For your information, is_stake indicates whether the match is Friendly or not. Some teams tend to not do their best on friendly matches, so is_stake handles these cases.

Similarly, I added is_worldcup to specially handle world cup matches.

```
match_df['rank_difference'] = match_df['home_team_fifa_rank'] - match_df['away_team_fifa_rank']
match_df['average_rank'] = (match_df['home_team_fifa_rank'] + match_df['away_team_fifa_rank'])/2
match_df['point_difference'] = match_df['home_team_total_fifa_points'] - match_df['away_team_total_fifa_points']
match_df['is_stake'] = match_df['tournament'] != 'Friendly'
match_df['is_worldcup'] = 'FIFA World Cup' in match_df['tournament']

match_df['score_difference'] = match_df['home_team_score'] - match_df['away_team_score'] # Note that this feature is not used in training
match_df['is_won'] = match_df['score_difference'] > 0 # Take draw as lost
```

Training

Let's try different machine learning models. In this notebook, we'll try the following.

1. Logistic Regression >>68.38
2. Support Vector Machines >>68.17
3. Gaussian Naive Bayes >>68.36

I'll use Logistic Regression for final prediction. Ensembling top 3 models may work better though.

```
# Logistic Regression

logreg = LogisticRegression()
logreg.fit(X_train, y_train)
lg_pred = logreg.predict(X_test)
acc_log = round(logreg.score(X_test, y_test) * 100, 2)
acc_log
```

68.38

```
# Support Vector Machines

svc = SVC()
svc.fit(X_train, y_train)
svm_pred = svc.predict(X_test)
acc_svc = round(svc.score(X_test, y_test) * 100, 2)
acc_svc
```

68.17

```
# Gaussian Naive Bayes

gaussian = GaussianNB()
gaussian.fit(X_train, y_train)
gnb_pred = gaussian.predict(X_test)
acc_gaussian = round(gaussian.score(X_test, y_test) * 100, 2)
acc_gaussian
```

68.36

Group Stage Match Prediction

Here in the prediction section for the group stage games through the data to see which team will collect the most points to win in the stage.

Result of game with all team

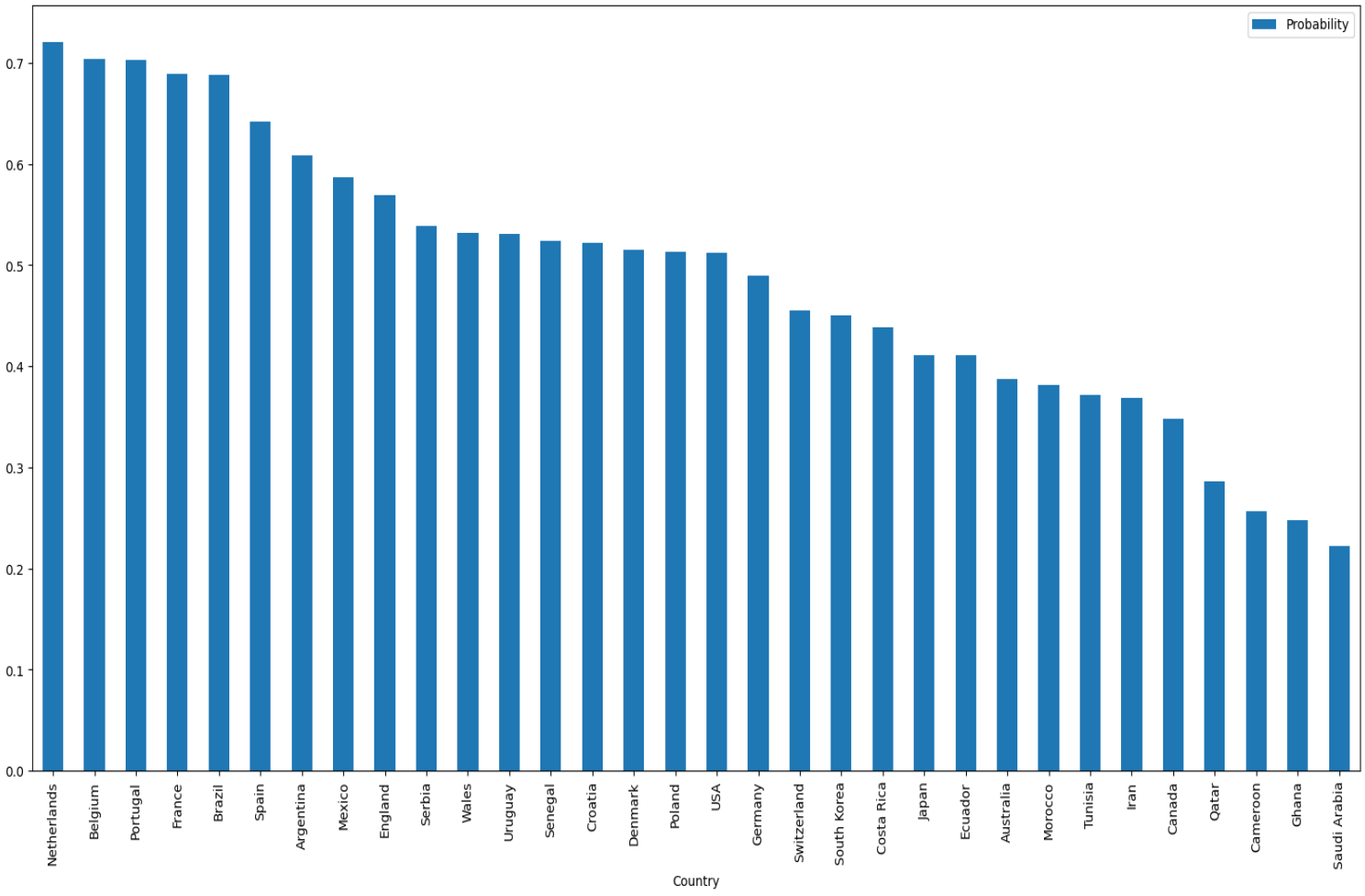
```
___Starting group C:___
Argentina vs. Saudi Arabia: Argentina wins with 0.69
Argentina vs. Mexico: Draw
Argentina vs. Poland: Argentina wins with 0.57
Saudi Arabia vs. Mexico: Mexico wins with 0.73
Saudi Arabia vs. Poland: Poland wins with 0.67
Mexico vs. Poland: Draw
___Starting group E:___
Germany vs. Japan: Draw
Germany vs. Spain: Spain wins with 0.58
Germany vs. Costa Rica: Germany wins with 0.55
Japan vs. Spain: Spain wins with 0.64
Japan vs. Costa Rica: Draw
Spain vs. Costa Rica: Spain wins with 0.58
___Starting group A:___
Senegal vs. Qatar: Senegal wins with 0.62
Senegal vs. Netherlands: Netherlands wins with 0.61
Senegal vs. Ecuador: Senegal wins with 0.59
Qatar vs. Netherlands: Netherlands wins with 0.74
Qatar vs. Ecuador: Ecuador wins with 0.57
Netherlands vs. Ecuador: Netherlands wins with 0.64
___Starting group F:___
Morocco vs. Croatia: Croatia wins with 0.60
Morocco vs. Belgium: Belgium wins with 0.66
Morocco vs. Canada: Morocco wins with 0.55
Croatia vs. Belgium: Belgium wins with 0.61
Croatia vs. Canada: Croatia wins with 0.60
Belgium vs. Canada: Belgium wins with 0.65
```

Point each team

```
___Starting group C:___
Argentina :  4.286353942719144
Saudi Arabia :  0.0
Mexico :  3.1963192020317854
Poland :  2.481271369033881
___Starting group E:___
Germany :  2.178141394621764
Japan :  0.9728076701631958
Spain :  5.380953594716832
Costa Rica :  0.511359888856254
___Starting group A:___
Senegal :  3.6079965458122643
Qatar :  0.0
Netherlands :  5.959668575161052
Ecuador :  1.7127215169764838
___Starting group F:___
Morocco :  1.652644981266103
Croatia :  3.6146314412505074
Belgium :  5.7678391556116795
Canada :  0.0
___Starting group D:___
Denmark :  3.4485978500610264
Tunisia :  0.4945877992220571
France :  5.696886533016041
Australia :  0.5054122007779429
___Starting group H:___
Uruguay :  2.5899996526460924
South Korea :  2.3520097476957256
Portugal :  5.816296001915481
Ghana :  0.0
___Starting group B:___
Iran :  0.0
England :  2.9227495013352045
USA :  2.6840236954486825
Wales :  2.6832559153963644
___Starting group G:___
Switzerland :  2.2706012878512096
Cameroon :  0.0
Brazil :  5.810936160477053
Serbia :  2.482411222170502
```

RESULT

In the end, by use data allowed us to predict which teams are more likely to qualify in group stage for fifa world cup 2022 round of 16.



LINK Github :

✓ Dadyar-sparky/fifa-word-cup (github.com)