National University of Computer

and Emerging Sciences

CS4048-Data Science

Project

Match Prediction

Group Members:

P19-0026-Shaheer Sarfraz

P19-0029-Hassam Ud Din

Project Description:

As the FIFA world cup, 2022 is an ongoing event that is captivating people's attention, predicting the winning team has always been the supporting fan desire. Considering this we have proposed a model to predict the winning team in the world cup using the domain knowledge of data science.

Technology:

- Python

Libraries:

- NumPy
- Pandas
- Matplotlib
- Seaborn
- Sklearn

Project Flow:

Taking into account the life cycle of data science, the project is divided into phases covering each step of the data science life cycle.

Question:

Which team is going to take the world cup this year and what are the attributes which positively and negatively affect a team's winning probability?

Acquire:

Dataset was Acquired from a GitHub repository as the dataset from Kaggle doesn't fit our requirements.

Link:https://github.com/jieguangzhou/FIFA-World-Cup-2022/blob/master/international_matches.csv

	date	home_team	away_team	home_team_continent	away_team_continent	home_team_fifa_rank	away_team_fifa_rank	home_team_total_fifa_points
0	1993-08-08	Bolivia	Uruguay	South America	South America	59	22	0
1	1993-08-08	Brazil	Mexico	South America	North America	8	14	0
2	1993-08-08	Ecuador	Venezuela	South America	South America	35	94	0
3	1993-08-08	Guinea	Sierra Leone	Africa	Africa	65	86	0
4	1993-08-08	Paraguay	Argentina	South America	South America	67	5	0
...	...	...	...	...	...	...	...	...
23916	2022-06-14	Moldova	Andorra	Europe	Europe	180	153	932
23917	2022-06-14	Liechtenstein	Latvia	Europe	Europe	192	135	895
23918	2022-06-14	Chile	Ghana	South America	Africa	28	60	1526
23919	2022-06-14	Japan	Tunisia	Asia	Africa	23	35	1553
23920	2022-06-14	Korea Republic	Egypt	Asia	Africa	29	32	1519

23921 rows × 25 columns

Figure 1 Preview of Dataset

ETL:

Since the dataset is taken from a single source and is in the form of a CSV file so there is no need for transformation.

Wrangling:

Initially, our dataset consisted of 23921 rows x 25 columns as we are working on only the semi-final teams the data for other teams is not for our use for which we have dropped other data to minimize our dataset for a better understanding and results.

Code:

```
df= df.query("home_team == 'Argentina' | home_team == 'France' | home_team == 'Morocco' | home_team == 'Croatia'")
```

Output:

	date	home_team	away_team	home_team_continent	away_team_continent	home_team_fifa_rank	away_team_fifa_rank	home_team_total_fifa_points
15	1993-08-22	Argentina	Peru	South America	South America	5	70	0
22	1993-08-29	Argentina	Paraguay	South America	South America	5	67	0
27	1993-09-05	Argentina	Colombia	South America	South America	5	19	0
47	1993-09-15	Morocco	Mali	Africa	Africa	33	74	0
68	1993-10-01	Morocco	Gabon	Africa	Africa	32	51	0
...	...	...	...	...	...	...	...	...
23741	2022-06-05	Argentina	Estonia	South America	Europe	4	110	1765
23753	2022-06-06	Croatia	France	Europe	Europe	16	3	1621
23804	2022-06-09	Morocco	South Africa	Africa	Africa	24	69	1551
23879	2022-06-13	Morocco	Liberia	Africa	Africa	24	149	1551
23885	2022-06-13	France	Croatia	Europe	Europe	3	16	1789

758 rows × 25 columns

Figure 2 Preview of the refined dataset

Now to get an overview of the dataset info() function is used to get an overview of Dtype and other attributes of data.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 758 entries, 15 to 23885
Data columns (total 28 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   date                            758 non-null    object
 1   home_team                       758 non-null    object
```

Figure 3 df.info()

Now the next step is to clean the missing values

```
df.isna().sum()
```

```
date                              0
home_team                         0
away_team                         0
home_team_continent               0
away_team_continent               0
home_team_fifa_rank               0
away_team_fifa_rank               0
home_team_total_fifa_points       0
away_team_total_fifa_points       0
home_team_score                   0
away_team_score                   0
tournament                        0
city                              0
country                           0
neutral_location                  0
shoot_out                         0
home_team_result                  0
home_team_goalkeeper_score      273
away_team_goalkeeper_score      377
home_team_mean_defense_score    273
home_team_mean_offense_score    273
home_team_mean_midfield_score   273
away_team_mean_defense_score    384
away_team_mean_offense_score    358
away_team_mean_midfield_score   370
dtype: int64
```

Figure 4 Preview of Null values

Filling null values with mean()

```
for x in wc_2022:
    for y in columns_contains_null:
        df[y].fillna(df[df[y[0:9]]==x][y].mean(),inplace = True)
```

Now after cleaning we can gain insights to predict the winner by analyzing different aspects of the teams.

Data Exploration:

- **Wins:**
  Team Streak is based on the team wins and loses.

```
Team_streak['team'].value_counts()
```

```
France       172
Morocco      155
Argentina    148
Croatia      113
Name: team, dtype: int64
```

Figure 5 Team Streak

- **Team Defense score:**
  The Team Defense score is based on the team's overall defense against other teams.

```
defence_scores
```

	Team	Df score
1	France	83.93
2	Argentina	83.35
3	Croatia	80.63
4	Morocco	79.59

Figure 6 Defense Score

- **Team Offense Score:**
  Team Offence score is based on the team's offense strength against other teams.

```
offence_scores
```

	Team	Of score
1	Argentina	88.40
2	France	87.39
3	Morocco	82.99
4	Croatia	82.83

Figure 7 Offense Score

Visualization:

Now as we have gathered the insights we can visualize our data.

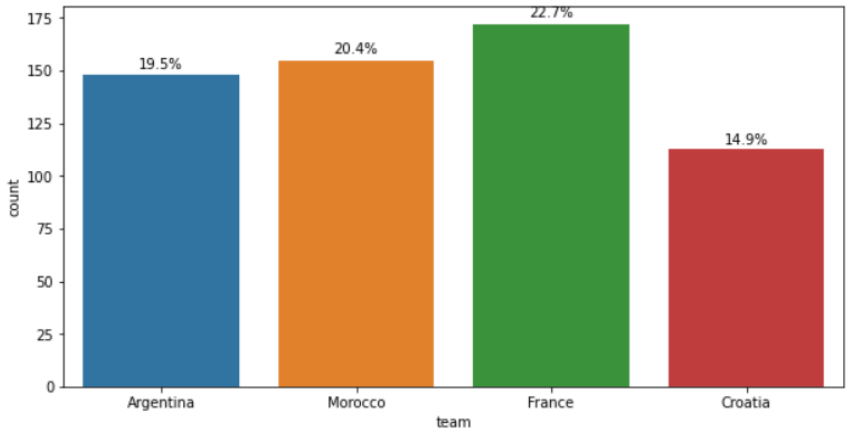

Figure 8 Preview of Team Wins

Figure 8 shows us the count plot of Team Streak which tells us which team has more wins.

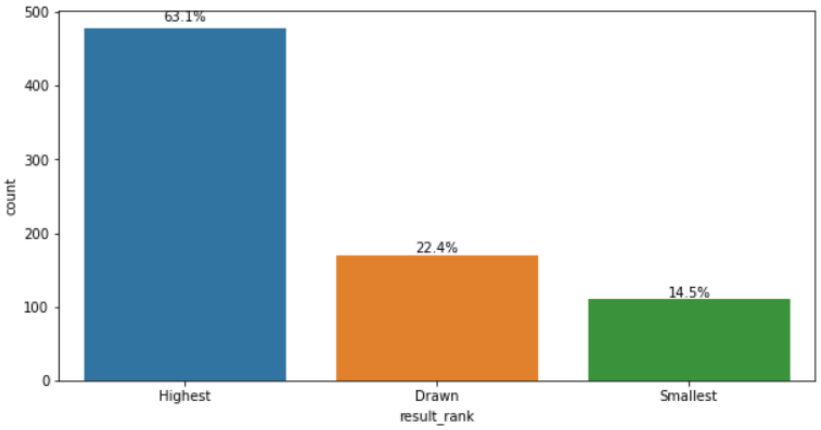

Figure 9 High-ranked vs Low-ranked

Visualizing the data for the highest-ranked teams against low-ranked teams which in Figure 9 shows us that when a highest-ranked team plays against a low-ranked team the chances of winning are highest as compared to losing or drawing.

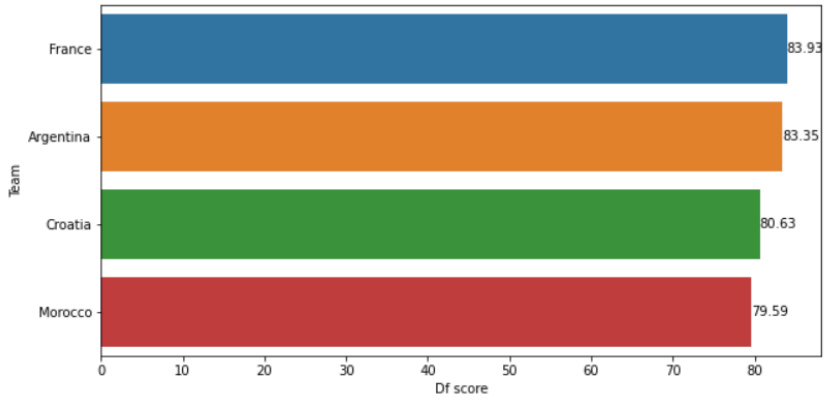

Figure 10 Defense score

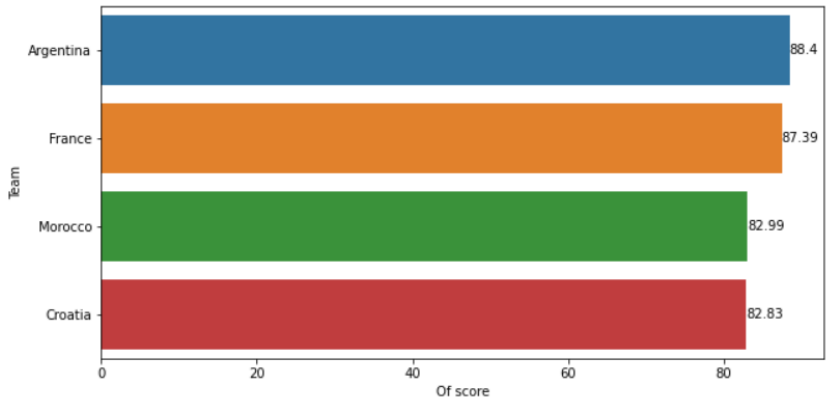

Figure 11 Offense score

Visualization of teams defense and offense score.

Results:

Considering different aspects of a team like defense score, offense score and team winning streak we can now predict the winning probability of a team.

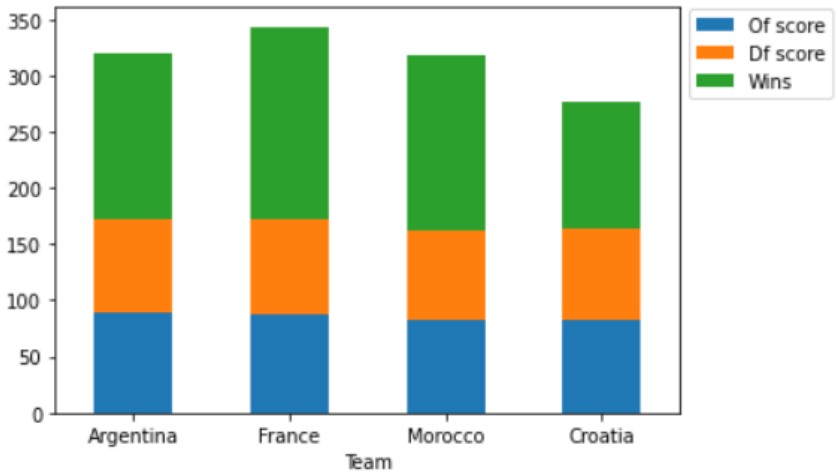

Figure 12 Results

Figure 12 shows us the visualization of results gained from the data of semi-finalist teams which shows us the probability of winning are higher for France considering some main factors which affect winning probability.

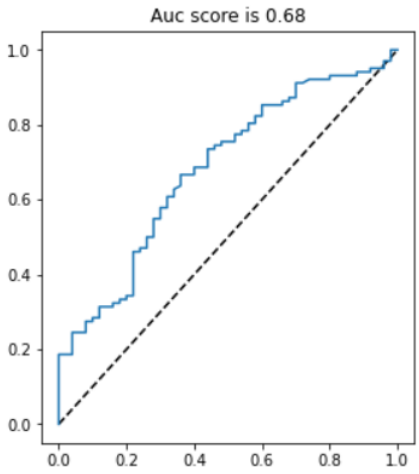

Figure 13 AUC score

Figure 13 shows us the AUC score of 0.68 which we got after training our data which is quite good. This shows that our above analysis which predicts the chances of winning for France is the most likely outcome.

Note: PDF contains only the code snippets main code is provided in the python notebook.