

# 2022 Soccer World Cup Prediction

Capstone Project


*Ancy Joseph - 07/02/23*



**FIFA WORLD CUP**  
**Qatar2022**



## AGENDA

1. *Project Background*
  2. *Data Analysis*
  3. *Feature Identification*
  4. *Modeling Results & Predictions*
  5. *Summary & Recommendations*
- 

# Project Background

## Context:

Children First Soccer (CFS) Ltd. (non-profit that financially supports underprivileged children to enter into the professional world of soccer) wants to promote soccer in the developing markets across Asia (e.g. Laos, Cambodia, Myanmar etc.) leading up to the 2026 FIFA World Cup in USA, Canada and Mexico. CFS would like to understand which competitive teams should drive promotional campaigns within these markets in order to create excitement and passion for the game. CFS believes this will set up the foundation for children (both boys & girls), specifically in the age group of 6-15, to be coached, trained and presumably selected for the 2027 FIFA U-17 Boys World Cup in China & Girls World Cup in New Zealand respectively.



## Criteria for Success:

CFS needs to understand the countries who have a likelihood to qualify into the Round of 16, Quarter Finals, Semi Finals and Finals specifically for the 2023 FIFA World Cup in Qatar in order to drive a 5 year strategy in the developing Asia markets leading up to the **U-17 World Cups in 2027**. The 5 year roadmap is segmented into 2 Phases:

**Phase 1:** Focused on creating excitement in the markets and;

**Phase 2:** Drive training programs across priority markets/ age groups.

This proposal is currently focused on Phase 1 only.

# Project Background (Cont..)

## Scope:

Prediction will be based only on teams who have participated in FIFA Soccer World Cups till date and their performance in international matches.

## Constraints:

Factors such as venue, host country weather, timing of the tournament, referee judgment, Video Assistant Referee (VAR) interventions, squad formation, in-game tactical switches, and player concentration and stamina all play a huge role in predicting the final outcome.

These elements are relatively new to sports science and unsure about how to apply them as influential statistical factors in an algorithm.

## Stakeholders:

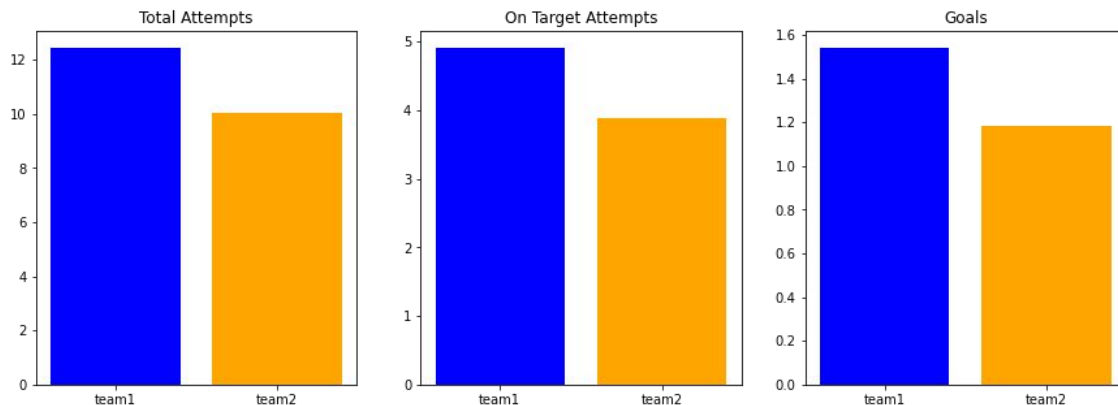
1. CFS CMO (Chief Marketing Officer)
2. CFS COO (Chief Operating Officer)
3. CFS APAC Regional Head
4. Ministry of Sports (APAC Developing Markets)

## Key Data Sources:

1. World Cups
2. 2022 World Cup Groups
3. 2022 World Cup Matches
4. World Cup Matches Stats
5. International Matches Stats
6. FIFA World Cup Ranking
7. FIFA International Matches

# Data Analysis

*Exploratory Data Analysis was conducted utilizing FIFA international matches data between 2012 and 2017. The limited data scope was selected to consider recency of players and teams who have actively participated in matches leading up to the World Cup. Team 1 are countries who play in their home stadiums and categorized as “Home Team” and Team 2 are countries who played away from their home stadiums and categorized as “Away Team”.*



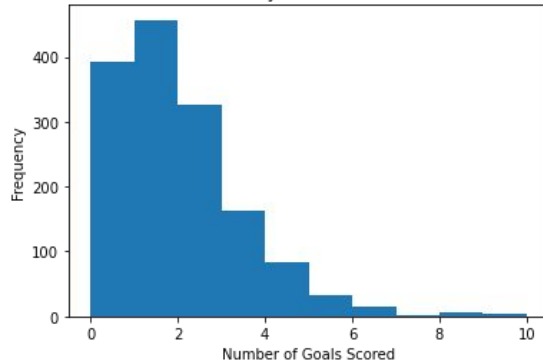
**Fig. 1**

**Comparison of Total Attempts, On Target Attempts & Goals scored (as shown in Fig.1):**

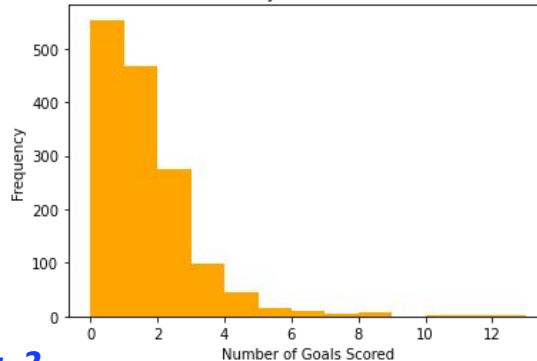
1. Team 1 (Home Team) has better results compared to Team 2 (Away Team)

# Data Analysis (Cont..)

Distribution of Goals Scored by team1 in matches from 2012 to 2017



Distribution of Goals Scored by team2 in matches from 2012 to 2017



**Fig. 2**

**Distribution of Goals scored (as shown in Fig.2):**

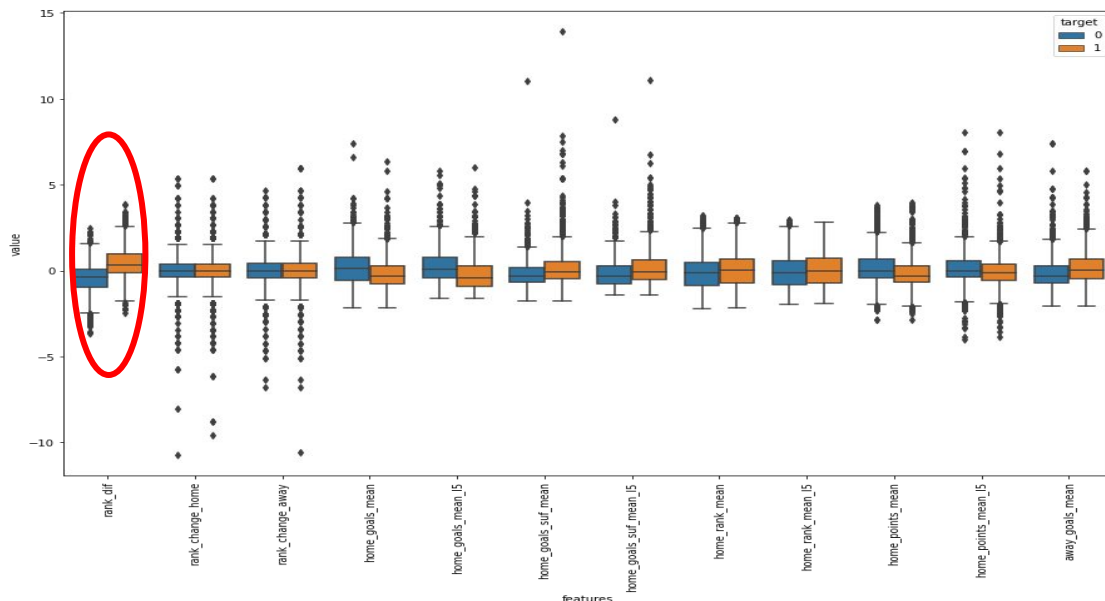
1. Team 2 (Away Team) has better goal distribution as compared to Team 1 (Home Team)

## CONCLUSION:

*It can be inferred that possession and attempts are crucial factors in determining the number of goals scored in a match, however, it is not the only factor, other factors like team strategies, player skills, and luck also play a role.*

# Features Identification

Identify the predictive features required to build the right machine learning model. This was conducted by comparing FIFA international match results and FIFA international ranking datasets. The 3 main categories of features that were identified are - Goals, Points & Rank.

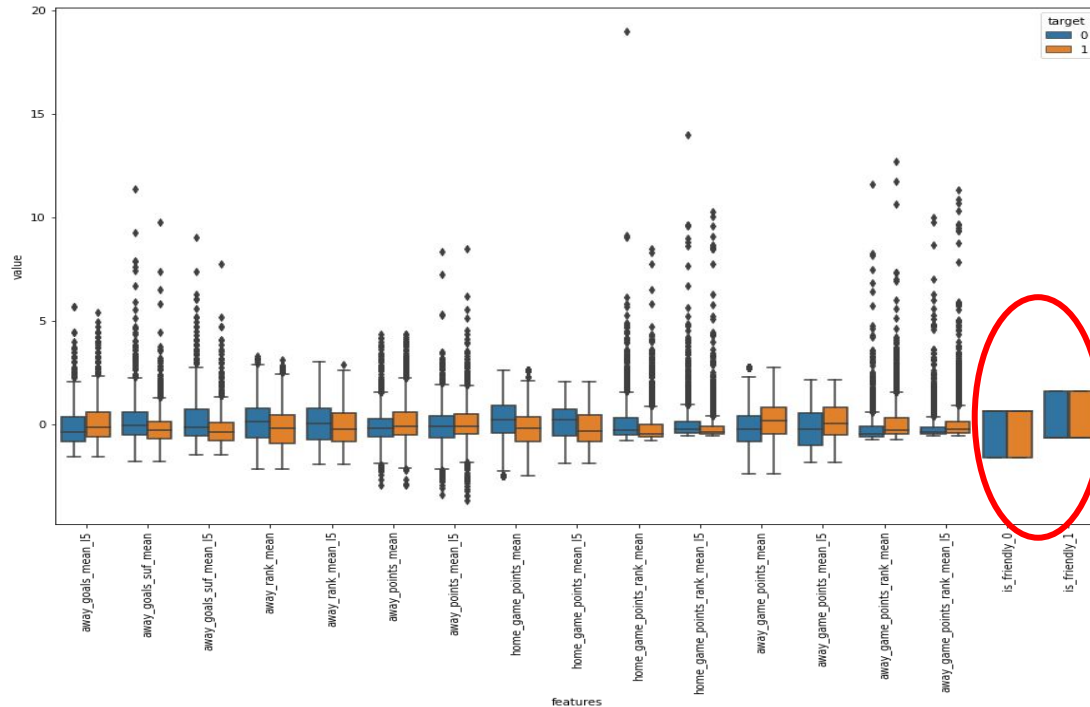


**Feature Analysis (as shown in Fig.3):**

1. “Rank difference” is a feature that calculates the difference between the FIFA rank of the home team compared to the away team and is considered as a good separator of data

**Fig. 3**

# Features Identification (Cont..)



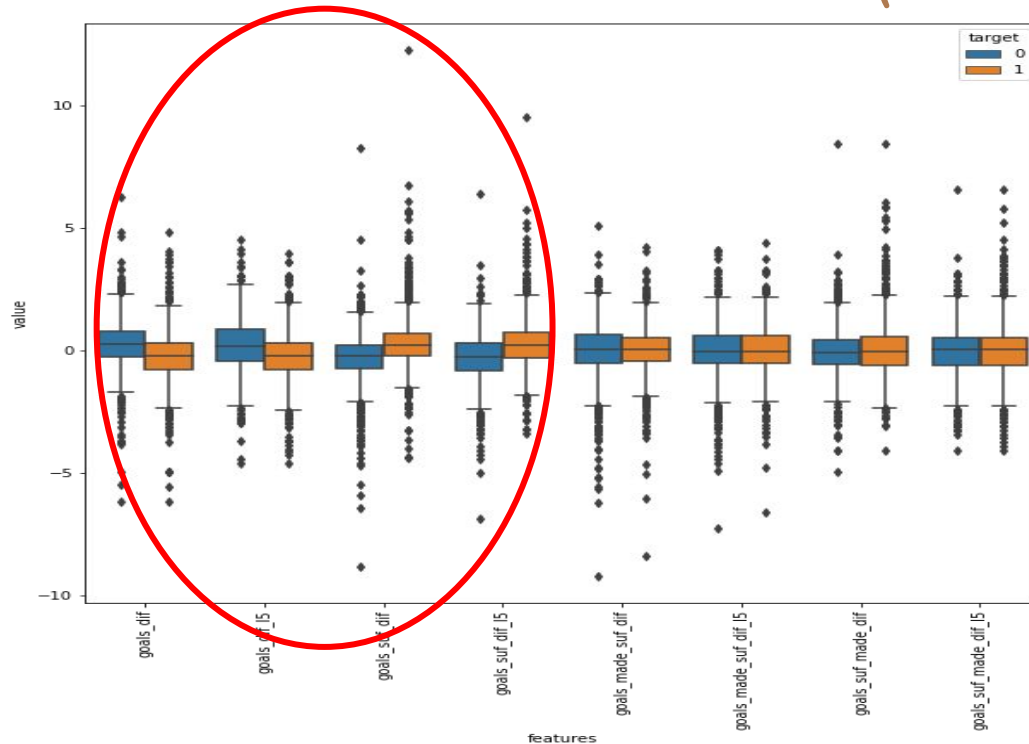
Feature Analysis (as shown in Fig.4):

1. “*Is Friendly*” is a feature that calculates if the game was an international friendly match or not and is considered as a good separator of data

Fig. 4



## Features Identification (Cont..)

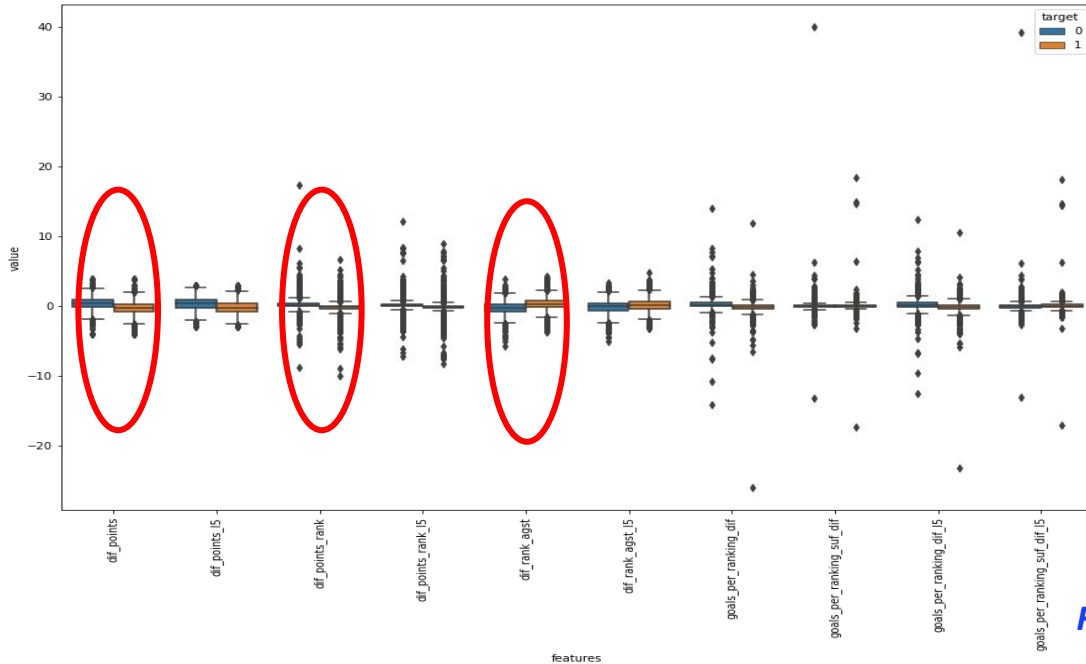


Feature Analysis (as shown in Fig.5):

1. “goal difference” and “goal suffered difference” are also considered as a good separators of data

Fig. 5

## Features Identification (Cont..)



Feature Analysis (as shown in Fig.6):

1. "difference of points" (full and last 5 games), "difference of points by ranking faced" (full and last 5 games) and "difference of rank faced" (full and last 5 games) are good features.

Fig. 6

# Features Identification (Cont..)

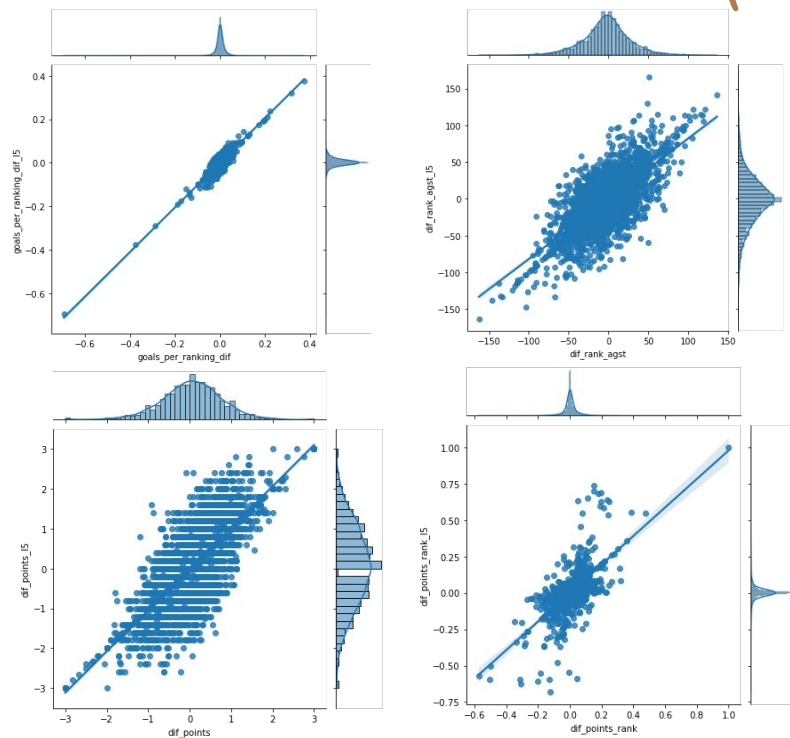


Fig. 7

## Feature Analysis (as shown in Fig.7):

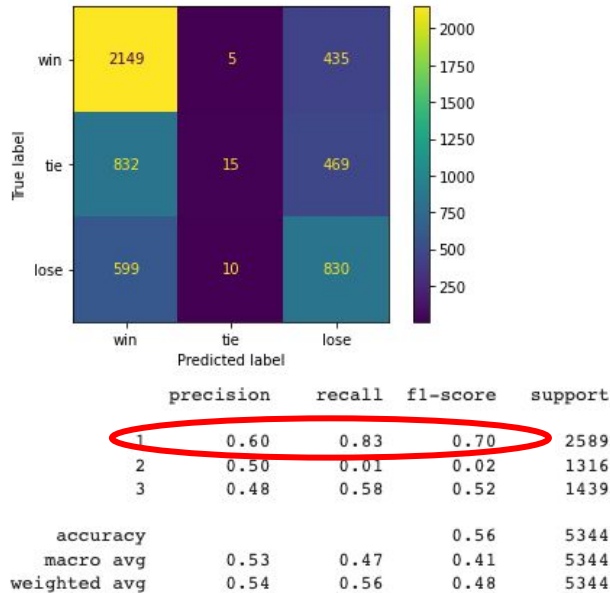
"Goals difference by ranking faced" and its last 5 games version has very similar distributions. So, we will use only the full version ( $\text{goals\_per\_ranking\_dif}$ ). For "differences of rank faced", "game points by rank faced" and "mean game points by rank faced", the two versions (full and 5 games) are not so similar. So, we decided to use both.

### Final features that were selected were:

1.  $\text{rank\_dif}$
2.  $\text{goals\_dif}$
3.  $\text{goals\_dif\_I5}$
4.  $\text{goals\_suf\_dif}$
5.  $\text{goals\_suf\_dif\_I5}$
6.  $\text{dif\_rank\_agst}$
7.  $\text{dif\_rank\_agst\_I5}$
8.  $\text{goals\_per\_ranking\_dif}$
9.  $\text{dif\_points\_rank}$
10.  $\text{dif\_points\_rank\_I5}$
11.  $\text{is\_friendly}$

# Modeling Results

11 predictive features were used to train and test 3 different machine learning models namely - Decision Tree, Logistic Regression & Random Forest models & identify the most accurate model.



**Modeling Results (as shown in Fig.8):**

1. Random Forest model was selected as the most accurate model to predict the 2022 FIFA Soccer World Cup Winner.

**Fig. 8**

# Predictions

*Random Forest model was utilized to predict the winners from group stages all the way through to the finals.*

	group	1	2
0	A	Netherlands	Qatar
1	B	England	Iran
2	C	Argentina	Mexico
3	D	France	Denmark
4	E	Germany	Spain
5	F	Belgium	Croatia
6	G	Brazil	Switzerland
7	H	Uruguay	South Korea

	home_team	away_team	home_pred	Winner	wm
Round of 16					
49	Netherlands	Iran	Win	Netherlands	W49
50	Argentina	Denmark	Win	Argentina	W50
51	England	Qatar	Win	England	W51
52	France	Mexico	Win	France	W52
53	Germany	Croatia	Win	Germany	W53
54	Brazil	South Korea	Win	Brazil	W54
55	Belgium	Spain	Lose	Spain	W55
56	Uruguay	Switzerland	Win	Uruguay	W56

# Predictions (Cont..)

		stage	ht	at	home_team	away_team	wm	home_pred	Winner		
Quarter finals											
57	Quarter-finals	W49	W50	Netherlands	Argentina	W57		Lose	Argentina		
58	Quarter-finals	W53	W54	Germany	Brazil	W58		Win	Germany		
59	Quarter-finals	W51	W52	England	France	W59		Win	England		
60	Quarter-finals	W55	W56	Spain	Uruguay	W60		Win	Spain		
		stage	ht	at	home_team	away_team	wm	home_pred	Winner	Losser	lm
Semi Final											
61	Semi-finals	W57	W58	Argentina	Germany	W61		Lose	Germany	Argentina	L61
62	Semi-finals	W59	W60	England	Spain	W62		Win	England	Spain	L62
		stage	ht	at	home_team	away_team	home_pred	Winner	wm		
Third match											
63	Third place	L61	L62	Argentina	Spain		Win	Argentina	W63		

# Predictions (Cont..)

	stage	ht	at	home_team	away_team	home_pred	Winner
Final							
64	Final	W61	W62	Germany	England	Win	Germany



# Summary & Conclusion

1. As per the predictive analysis completed and to meet the Phase 1 requirements of the proposal, the recommendation is for CFS Ltd. to focus their promotional campaigns on the **Top 14 countries** that were predicted to reach the Round of 16, with a special emphasis on **South Korea** as they were the only Asian country predicted to make it to the next round.
2. This should also be supported by highlighting marquee players like **Lionel Messi, Cristiano Ronaldo** as well as young and upcoming talent like **Kylian Mbappe, Cody Gakpo etc.** and their corresponding league and international impact (statistics) to create additional excitement and fan following.

## **Additional observations:**

Additional insights focused on Asian datasets e.g. *AFC (Asian Football Confederation) Cup stats, Asian players stats from European leagues* etc. will definitely boost the participation rate in the focused markets however these datasets were not available in the public domain for analysis & deeper insights.



IN REALITY....  
IT WAS! :)





Q&A



**THANK YOU!**

