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 Analysis*
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".

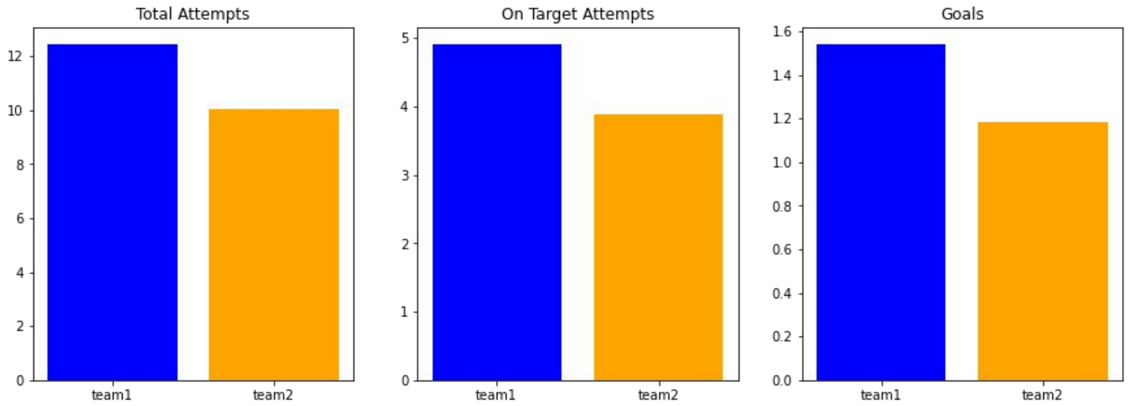


Fig. 1

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..)

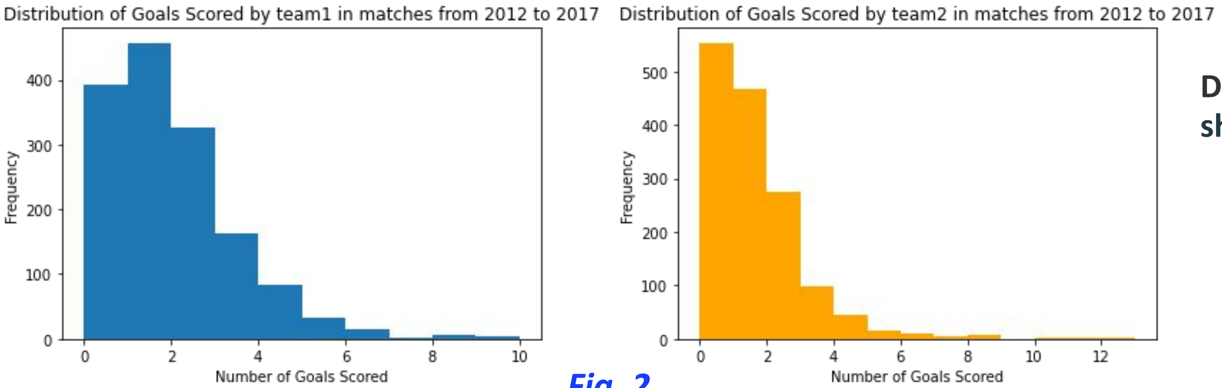


Fig. 2

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:

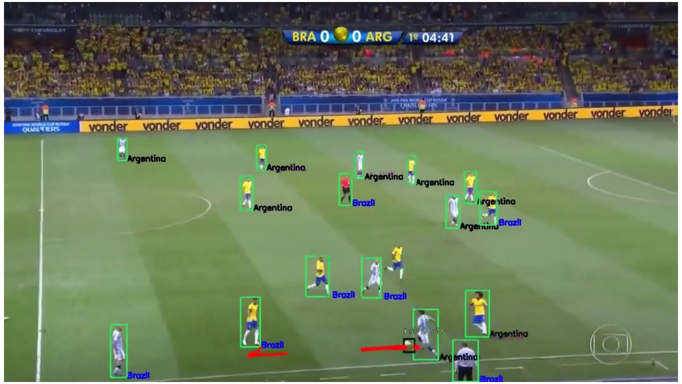
It can be inferred that possession and attempts are crucial factors in determining the number of goals scored in a match, however, it is not the only factor, other factors like team strategies, player skills, and luck also play a role.

Feature Significance

Data Scientists who have analyzed the game for many years have typically focused on *__3 broad feature categories__* namely: Goals, Points & Rank. Within these categories, there are many different variables that have to be considered in order to identify further nuances on how each feature influences the outcome of a game. Some examples used in this study are:

1. **Goals** - # of goals scored by the team, scored against by other teams, scored in home games, scored in away games etc.
2. **Points** - # of points won in home games, # of points won in away games etc.
3. **Rank** - FIFA Ranking

In the modern era, Data Scientists have access to even more granular data such as heat map of players during the game, effectiveness of formations etc. which I've not considered for this analysis as these datasets were not available in the public domain.



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.

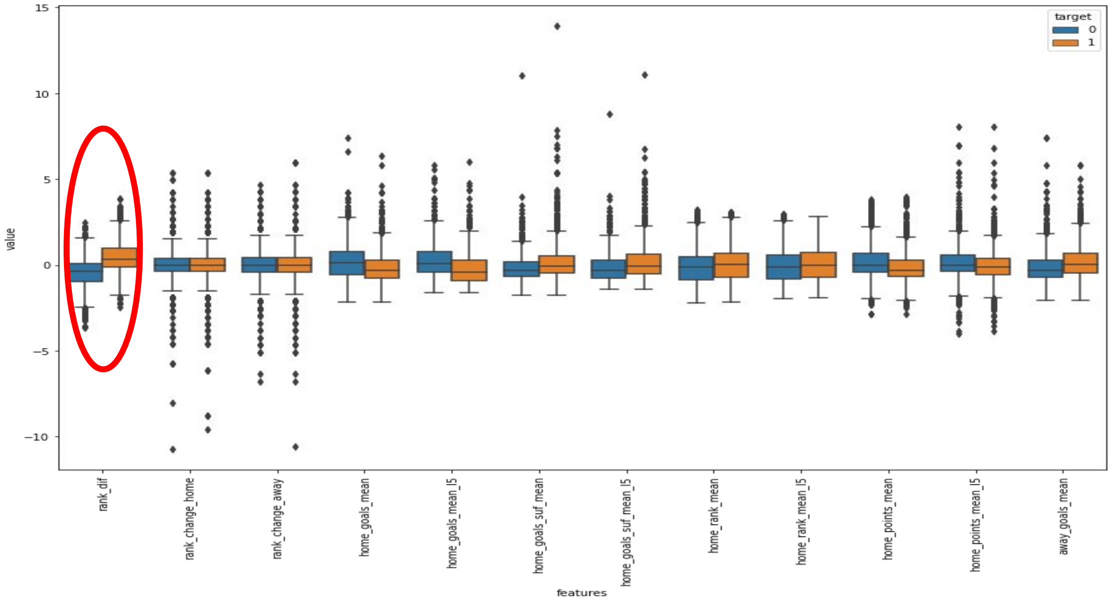


Fig. 3

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

Features Identification (Cont..)

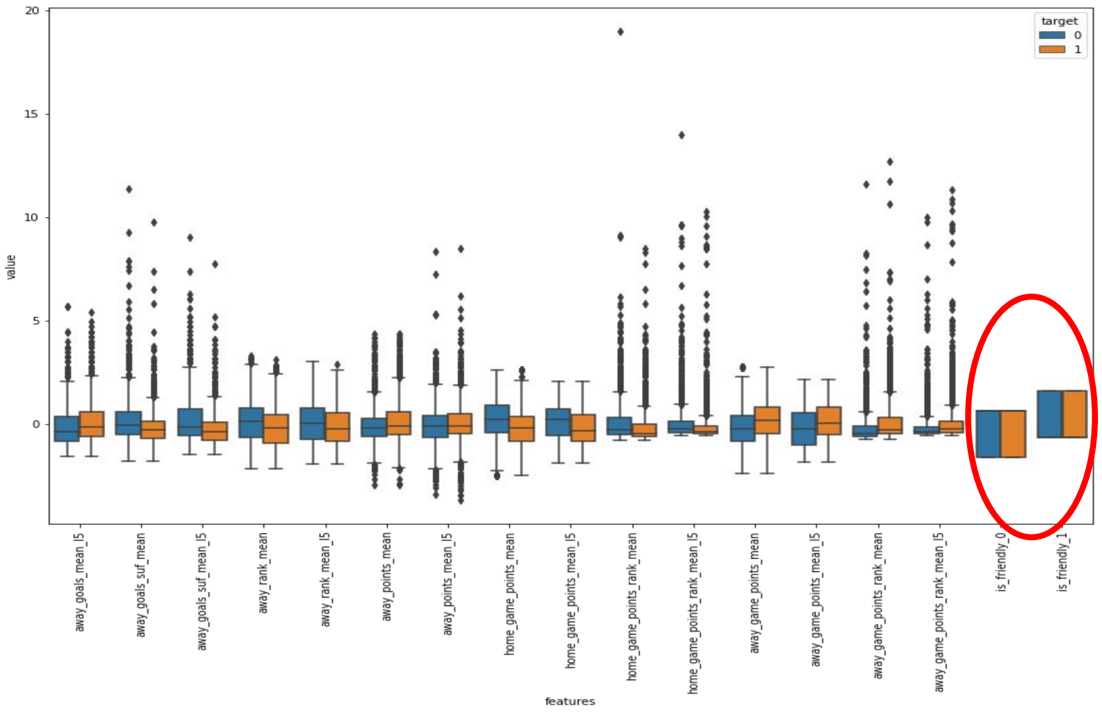


Fig. 4

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

Features Identification (Cont..)

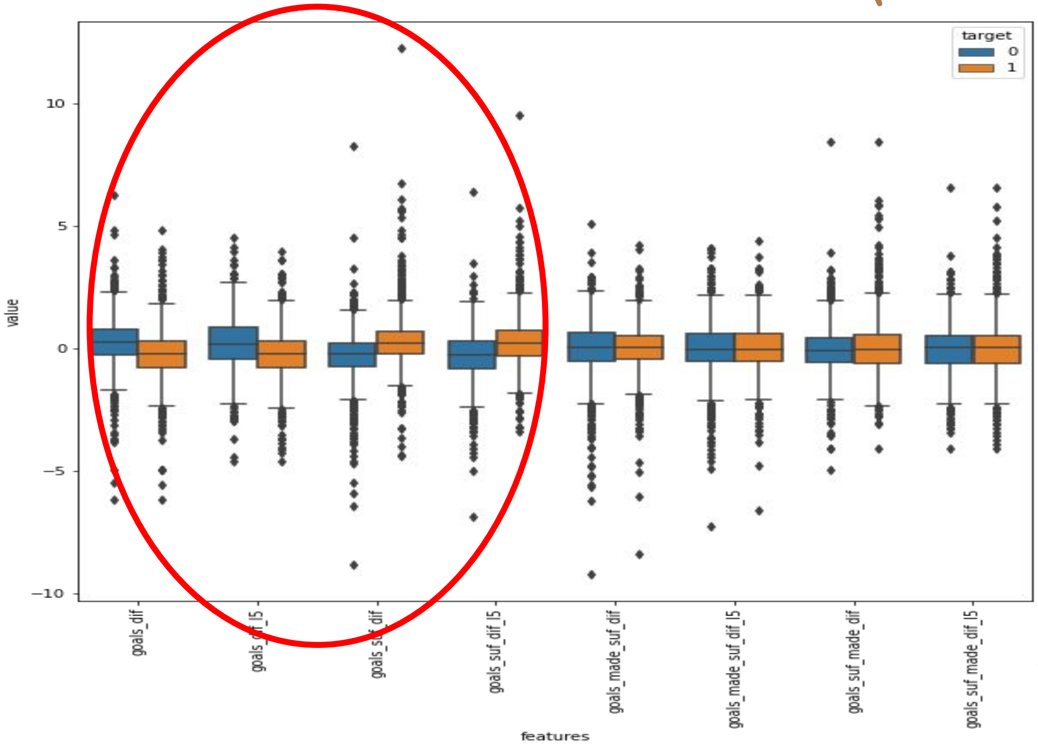


Fig. 5

Feature Analysis (as shown in Fig.5):

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

Features Identification (Cont..)

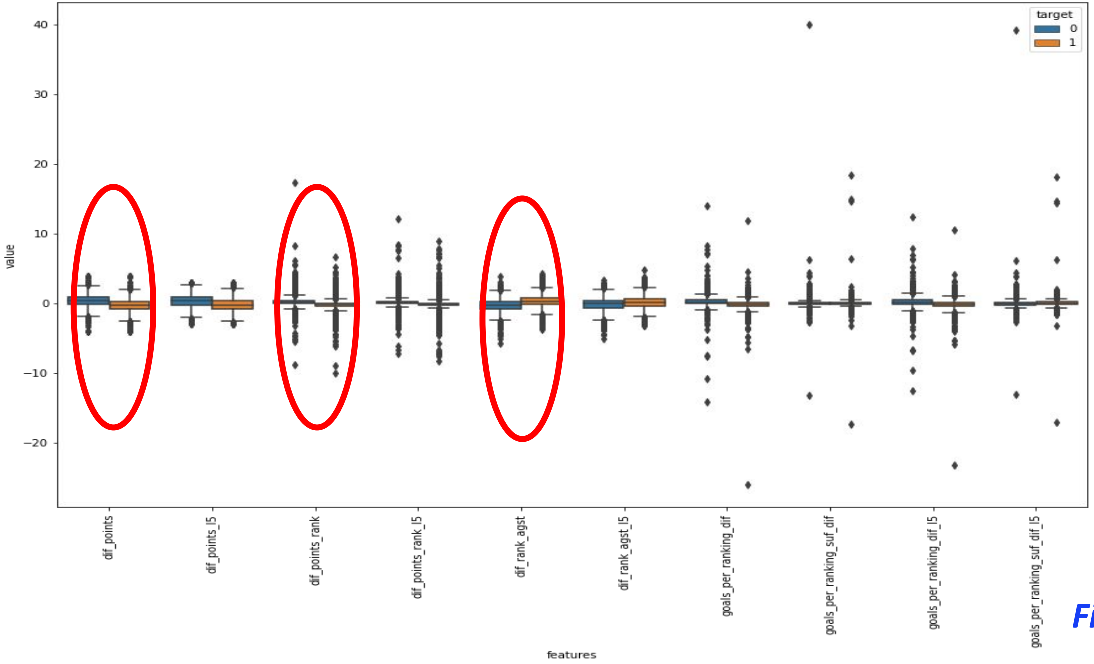


Fig. 6

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.

Features Identification (Cont..)

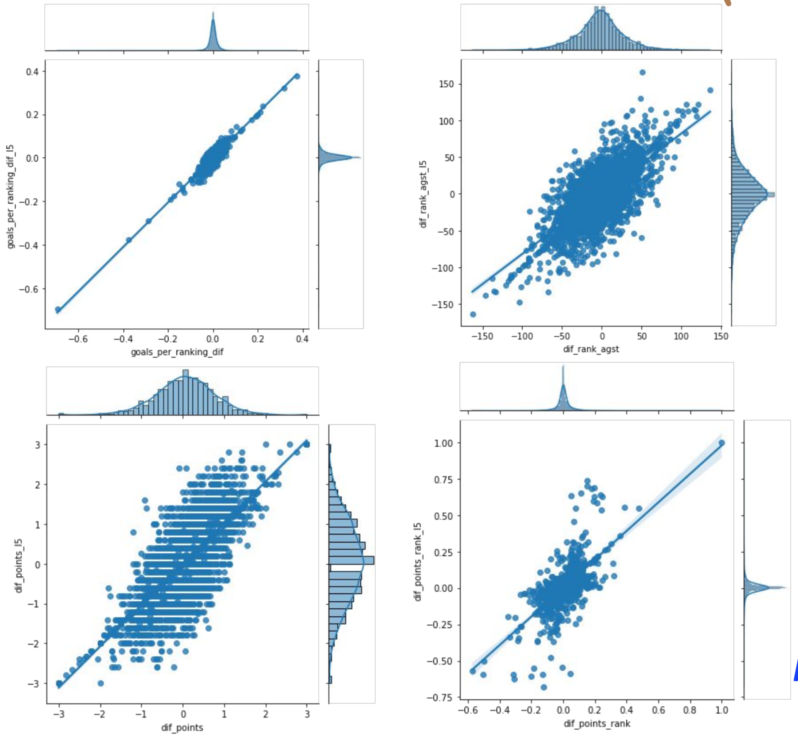


Fig. 7

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 (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. rank_dif
2. goals_dif
3. goals_dif_l5
4. goals_suf_dif
5. goals_suf_dif_l5
6. dif_rank_agst
7. dif_rank_agst_l5
8. goals_per_ranking_dif
9. dif_points_rank
10. dif_points_rank_l5
11. is_friendly_0
12. is_friendly_1

Feature Importance

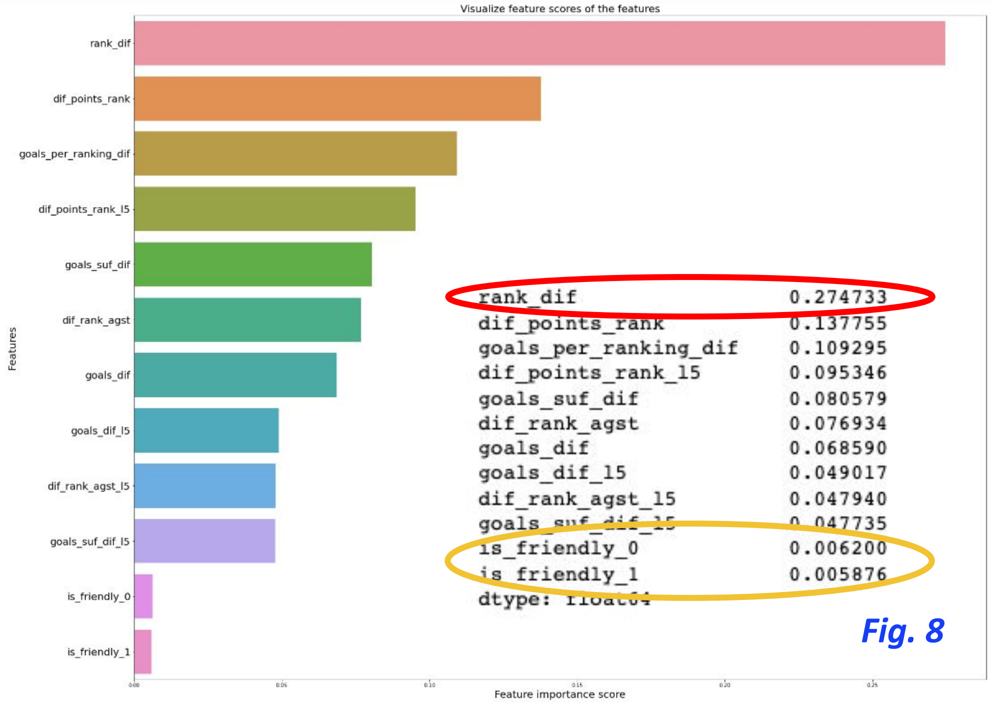


Fig. 8

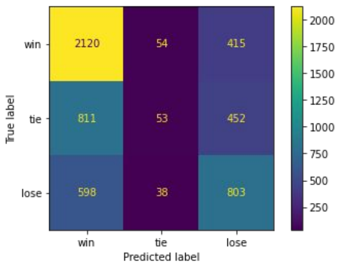
Feature Importance (as shown in Fig.8):

1. "*rank_dif*" is the most important feature and both "*is_friendly*" features are the lowest in importance.
2. Since friendly games leading up to a world cup tournament is pretty significant from a match preparation and tournament readiness standpoint, both "*is_friendly*" features were included in the predictive model selection process.

Modeling Results

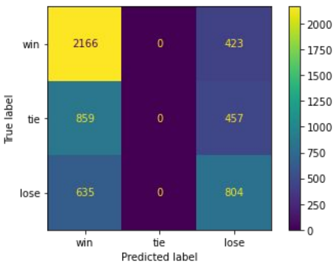
***12 predictive features** were used to train and test <u>3 different machine learning models</u> to identify the most appropriate model.*

1. Decision Tree



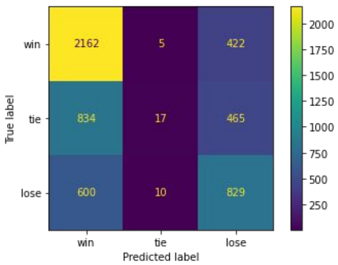
	precision	recall	f1-score	support
1	0.60	0.82	0.69	2589
2	0.37	0.04	0.07	1316
3	0.48	0.56	0.52	1439
accuracy			0.56	5344
macro avg	0.48	0.47	0.43	5344
weighted avg	0.51	0.56	0.49	5344

2. Logistic Regression



	precision	recall	f1-score	support
1	0.59	0.84	0.69	2589
2	0.00	0.00	0.00	1316
3	0.48	0.56	0.51	1439
accuracy			0.56	5344
macro avg	0.36	0.47	0.40	5344
weighted avg	0.42	0.56	0.47	5344

3. Random Forest



	precision	recall	f1-score	support
1	0.60	0.84	0.70	2589
2	0.53	0.01	0.03	1316
3	0.48	0.58	0.53	1439
accuracy			0.56	5344
macro avg	0.54	0.47	0.42	5344
weighted avg	0.55	0.56	0.49	5344

Fig. 9

Modeling Results (as shown in Fig.9):

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

Predictions - *Group Stages*

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

Model Prediction

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

In Reality

Netherlands	Senegal
England	United States
Argentina	Poland
France	Australia
Japan	Spain
Morocco	Croatia
Brazil	Switzerland
Portugal	South Korea

Predictions - *Round of 16*

Random Forest model was utilized to predict the winners from Round of 16!

Model Prediction

	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	Spain	Croatia	Win	Spain	W53
54	Brazil	South Korea	Win	Brazil	W54
55	Belgium	Germany	Lose	Germany	W55
56	Uruguay	Switzerland	Win	Uruguay	W56

In Reality

- ✅ Netherlands
- ✅ Argentina
- ✅ England
- ✅ France
- ❌ Croatia
- ✅ Brazil
- ❌ Morocco
- ❌ Portugal

Predictions - *Quarter Finals*

Random Forest model was utilized to predict the winners from Quarter Finals!

Model Prediction

	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	Spain	Brazil	W58	Lose	Brazil
59	Quarter-finals	W51	W52	England	France	W59	Win	England
60	Quarter-finals	W55	W56	Germany	Uruguay	W60	Win	Germany

In Reality

- ✅ Argentina
- ❌ Croatia
- ❌ France
- ❌ Morocco

Predictions - *Semi Finals*

Random Forest model was utilized to predict the winners from Semi Finals!

Model Prediction

	stage	ht	at	home_team	away_team	wm	home_pred	Winner	Losser	lm
Semi Final										
61	Semi-finals	W57	W58	Argentina	Brazil	W61	Win	Argentina	Brazil	L61
62	Semi-finals	W59	W60	England	Germany	W62	Win	England	Germany	L62

In Reality

✅ Argentina

❌ France

Predictions - *Third Place Match*

Random Forest model was utilized to predict the winners from Third Place Match!

Model Prediction

	stage	ht	at	home_team	away_team	home_pred	Winner	wm
Third match								
63	Third place	L61	L62	Brazil	Germany	Win	Brazil	W63

In Reality

Croatia

Predictions - *Finals*

Random Forest model was utilized to predict the winners from Finals!

Model Prediction

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

In Reality

Argentina

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.

*“ Thanks to **Daniel Wu** for his relentless support and thought leadership all throughout the project as my Springboard mentor, **Kenneth Gil-Pasquel** for troubleshooting and resolving my GitHub, Jupyter notebook & Google collab queries and **DJ Sarkar** for quick and timely responses to subject matter related doubts. ”*

Q&A

THANK YOU!