DATASOC X ATLASSIAN

DATATHON

JAKE
WARBY

WILLIAM
LI

TEAM ACASO

2021

FIFA WORLD CUP
Qatar2022

EXECUTIVE SUMMARY

How can FIFA predict the 32 qualifying teams for the 2022 FIFA World Cup?

Project Tasks	Prediction of 32 qualifying teams
Problem Analysis	• Countless possibilities with Confederation qualifying game lineups • Modelling features selection and final model choice
Results	A highly robust model which simulated 2018 Cup results with **75%** accuracy

TASK 1 - PREDICTING THE FINAL 32

Our model predicts the following 32 teams...

Qatar

CONCACAF

Mexico

U.S.A

Canada

Costa Rica

AFC

Iran

Australia

South Korea

Japan

Saudi Arabia

CONMEBOL

Brazil

Argentina

Uruguay

Colombia

CAF

Senegal

Nigeria

Tunisia

Morocco

Egypt

UEFA

France

Belgium

England

Germany

Denmark

Portugal

Spain

Italy

Croatia

Netherlands

Switzerland

Wales

Sweden

Methodology

1

Data cleaning

Data is collected and pre-processed

2

Model Building

A robust model is created to predict outcomes of games

3

Simulations

Qualification outcomes are simulated

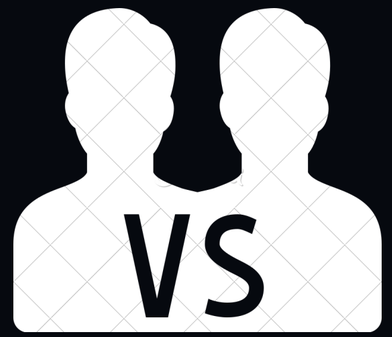
Model Features

Difference in FIFA National team Rating

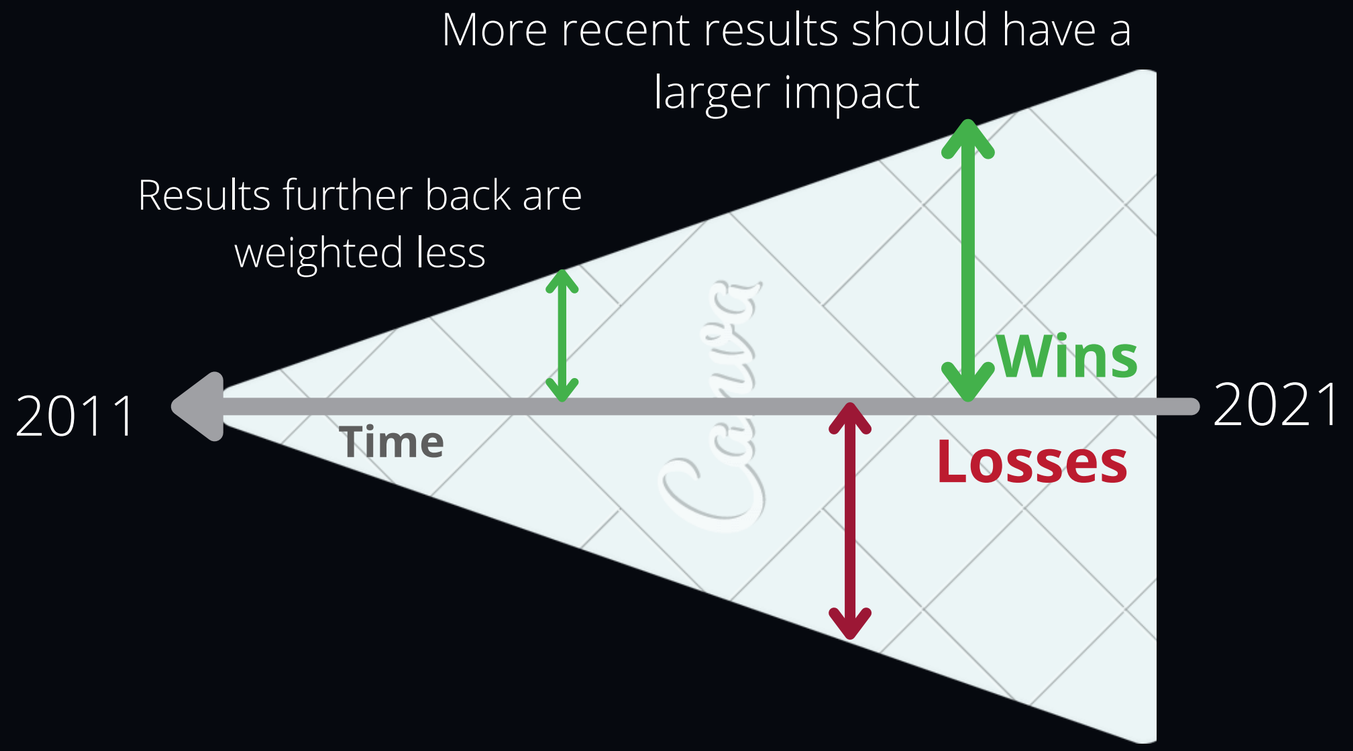
RK		Team	PTS	+/-
1		BEL	1832.33	9.99
2		BRA	1811.73	14.06
3		ENG	1755.44 ▲	2.61
4		FRA	1754.31 ▼	-7.46
5		ITA	1735.73	-8.94

Difference in Average FIFA Ultimate Team rating for top 15 players of each country

Recency-adjusted Features for head to head matches and regular matches



Features accounting for recency are superior



Recency score =

$$\Sigma \mathbb{1}_{\{Outcome = Home\}} \times \theta - \Sigma \mathbb{1}_{\{Outcome = Home\}} \times \theta$$

$$\theta = 1 - \frac{\#\ Days}{n \times 365}$$

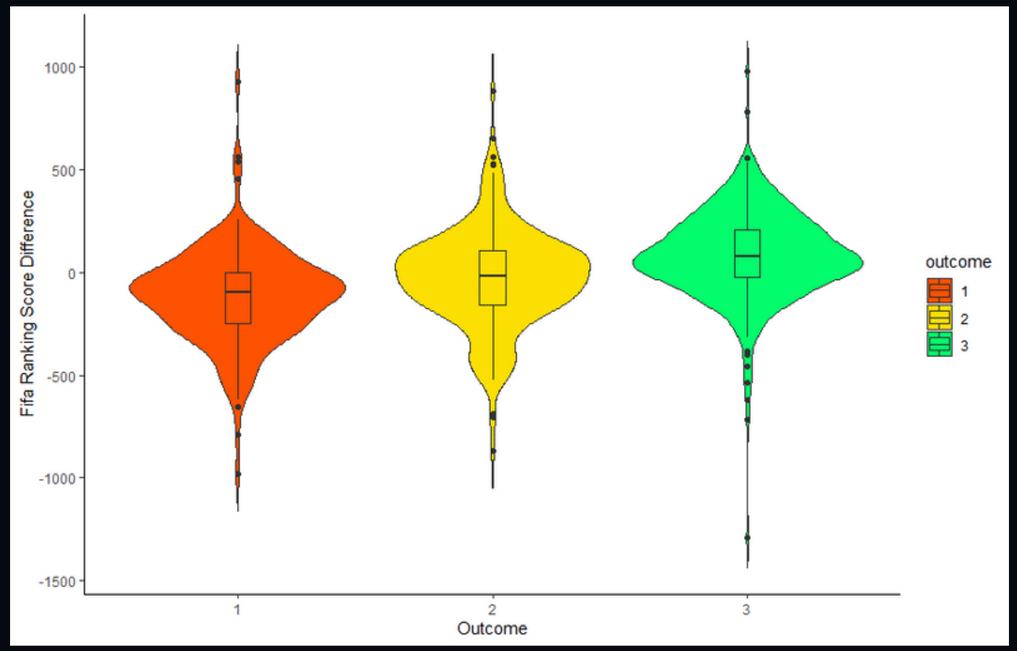
n = decay duration in years

e.g. on diagram, left: n = 10

Features that account for recency have stronger predictive power than their unweighted counterparts

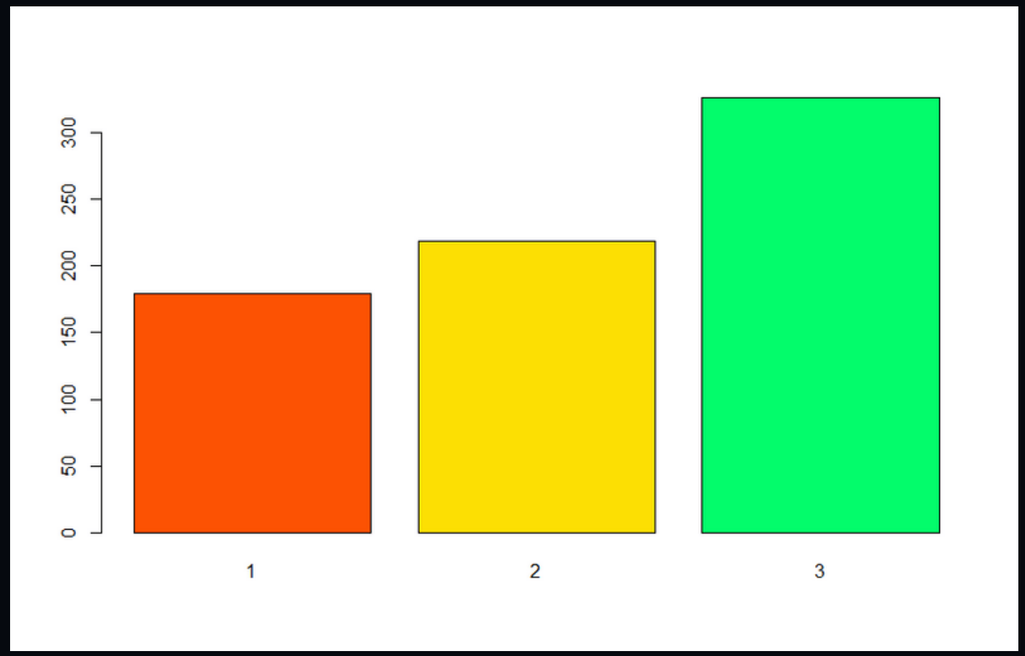
Notes on Features

Trend in Modes and Medians.
Increasing likelihood of home team winning with higher differentials.



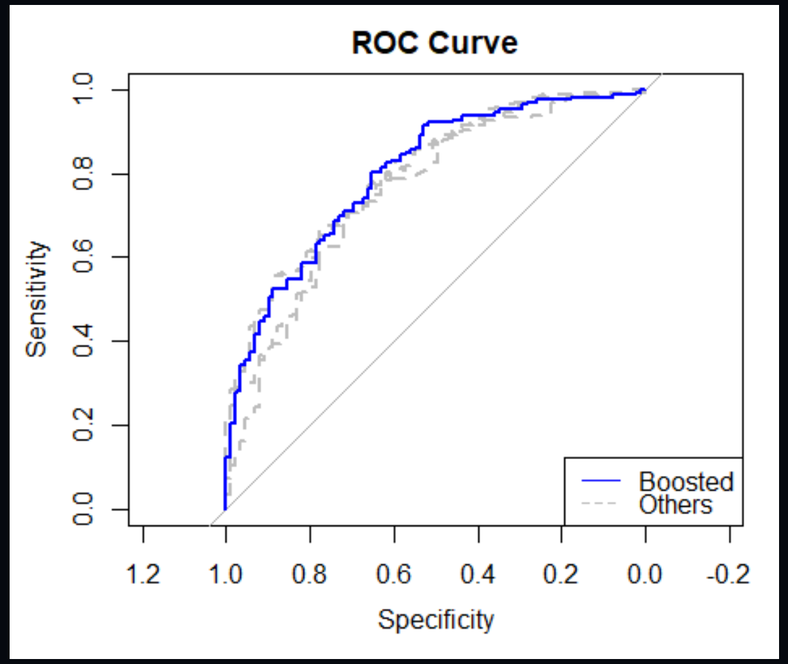
1 = away win 2 = draw 3 = home win

To Account for home team advantage, all features are differentials in respect to the home team



Model Selection

After 1000 repeats of 10 fold cross-validation over a range of models, the boosted model was determined to be the best.

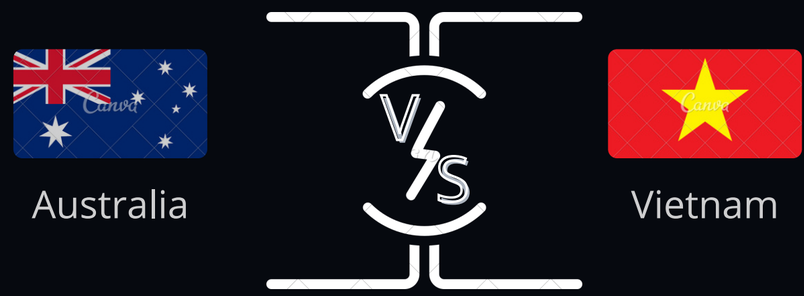


Model	Hyperparameters	logLoss	AUC	Accuracy
Boosted Trees	n.trees = 200, depth = 1	0.9788596	0.6777643	0.5308239

Simulations

Initial Scoreboard

Vietnam	6
Saudi Arabia	6
Australia	4
Japan	0



$$P_{qual} = \frac{\#SimQualification}{\#Sims}$$

Updated Scoreboard

Australia	7
Vietnam	6
Saudi Arabia	6
Japan	0

Evaluation using 2018 FIFA World Cup

Naive prediction

Predict the 32 teams from 2014 to all qualify for 2018

20 out of 32 Correct

Basic prediction

For every future match, predict the team with higher FIFA rating to win

RK	Team
1	BEL
2	BRA
3	ENG

23 out of 32 Correct

Our model

Gradient boosting model with 4 features, followed by simulations

24 out of 32 Correct

THANK YOU

FOR WATCHING

APPENDIX

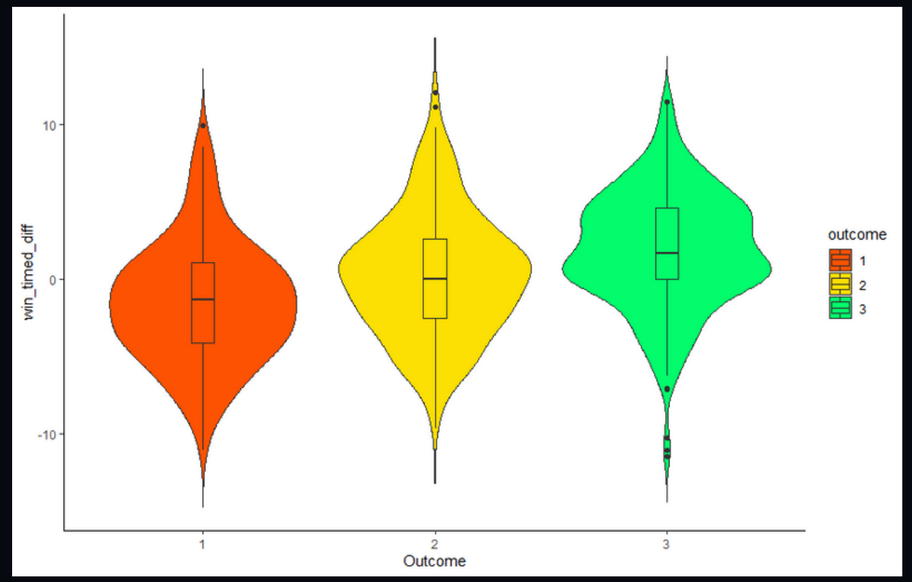
RECENCY IMPROVEMENT

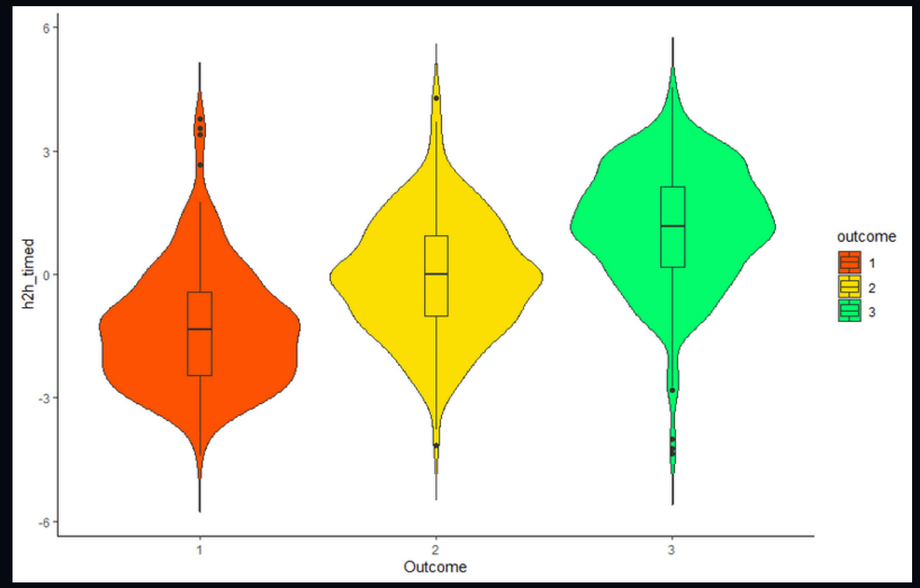
APPENDIX 1

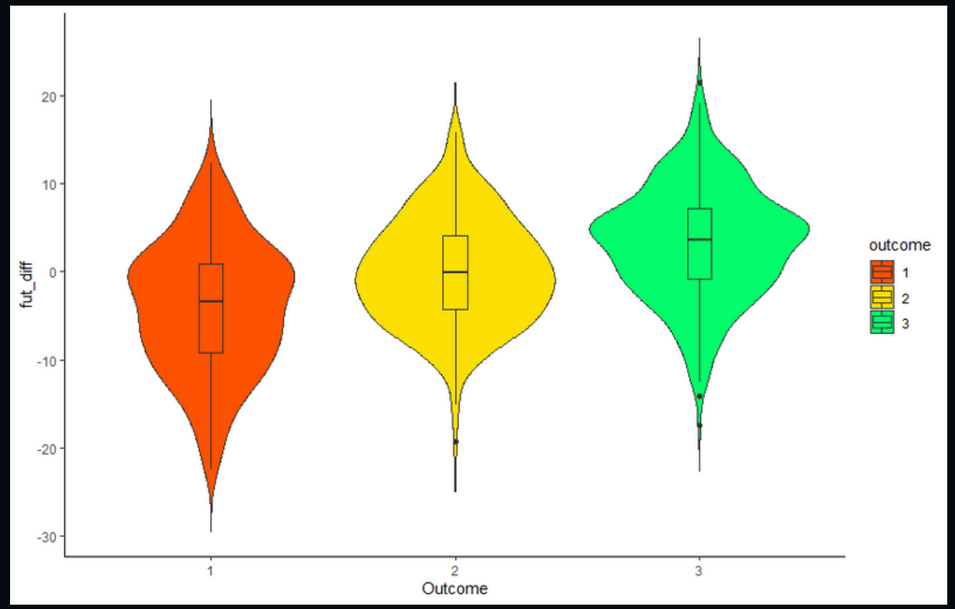
Model	logLoss	AUC	Accuracy
No Recency Adjustment	0.9879501	0.6767787	0.526508
Recency Adjustment	0.9788596	0.6777643	0.5308239

PREDICTOR V RESPONSE GRAPHS

APPENDIX 2



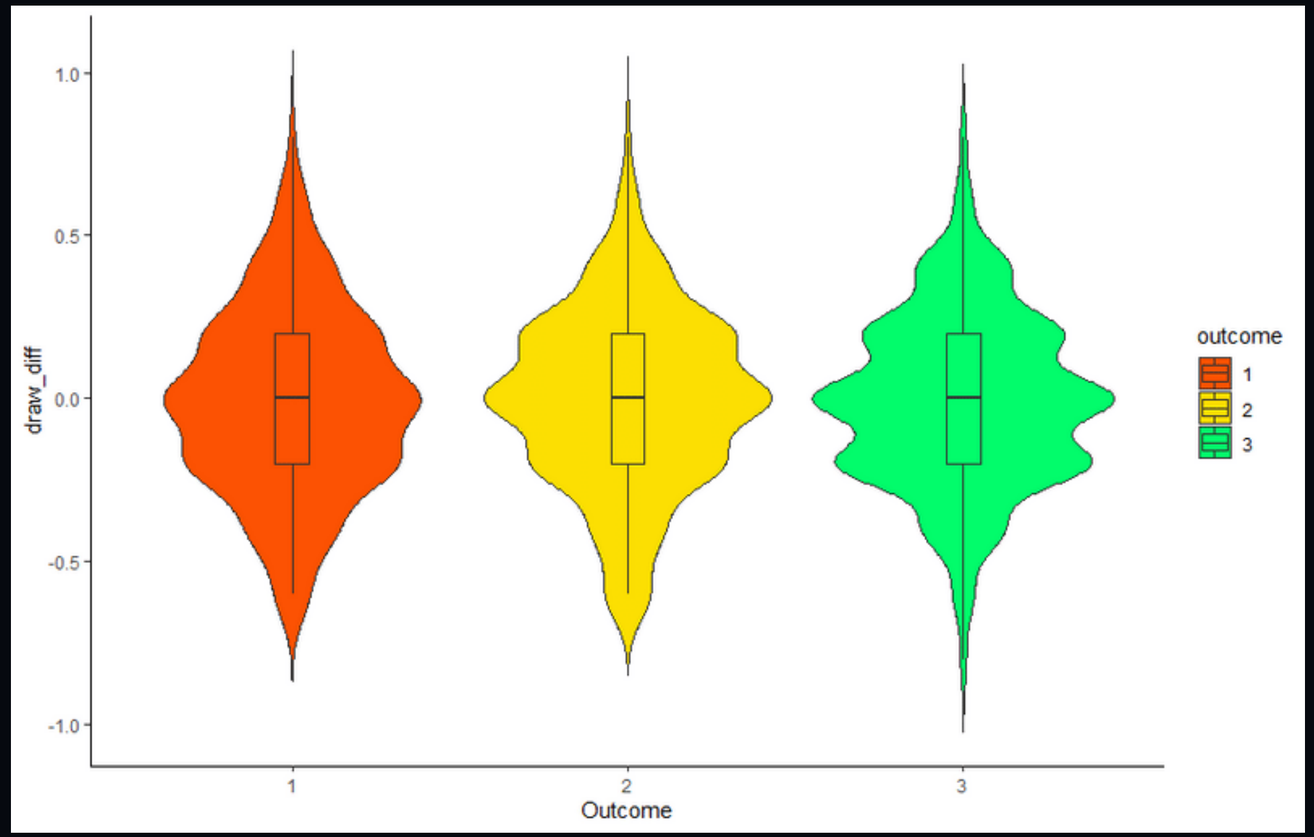




FIFA WORLD CUP
Qatar2022

APPENDIX 3

Rejected Feature



MODELLING RESULTS

APPENDIX 4

Model Results:

Model	Hyperparameters	logLoss	AUC	Accuracy
Lda	N/A	0.9847766	0.6618722	0.5229828
Qda	N/A	1.011474	0.660501	0.5192116
Multinomial	N/A	0.9842495	0.6681238	0.5220414
Random Forest	mtry = 1, min.node = 10	1.006935	0.6618912	0.5091618
Boosted Trees	n.trees = 200, depth = 1	0.9788596	0.6777643	0.5308239

Validation Type
K-fold CV
Folds
10
Repeats
100

FIFA WORLD CUP Qatar2022

MODELS USED

```
model_1 = gbm(outcome~h2h_timed + win_timed_diff + fifa_diff+fut_diff, data = na.omit(update_frame2[,c(12,17,19,21,22)]), n.trees = 200)
model_2 = gbm(outcome~win_timed_diff + fifa_diff + fut_diff, data = na.omit(update_frame2[,c(17,19,21,22)]), n.trees = 200)
model_3 = gbm(outcome~h2h_timed+win_timed_diff+fifa_diff,data=na.omit(update_frame2[,c(12,17,19,21)]), n.trees = 200)
model_4 = gbm(outcome~win_timed_diff+fifa_diff,data=na.omit(update_frame2[,c(17,19,21)]), n.trees = 200)
```

Final 32 Teams Predicted for 2022 and their qualifying probabilities

Team	Quals
Mexico	1
United States	0.9787
Canada	0.5609
Brazil	1
Argentina	0.9987
Uruguay	0.9908
Colombia	0.8567
Iran	0.9874
South Korea	0.8126
Australia	0.8129
Japan	0.7406
Senegal	0.9824
Nigeria	0.9583
Tunisia	0.9373
Morocco	0.8689

Team	Quals
Egypt	0.6614
France	1
Belgium	1
England	1
Germany	0.9989
Denmark	0.9983
Portugal	0.9977
Spain	0.9811
Italy	0.974
Croatia	0.941
Netherlands	0.9246
Switzerland	0.8805
Wales	0.5484
Sweden	0.5198
Saudi Arabia	0.7292
Costa Rica	0.4226

2018 FIFA World Cup Qualification

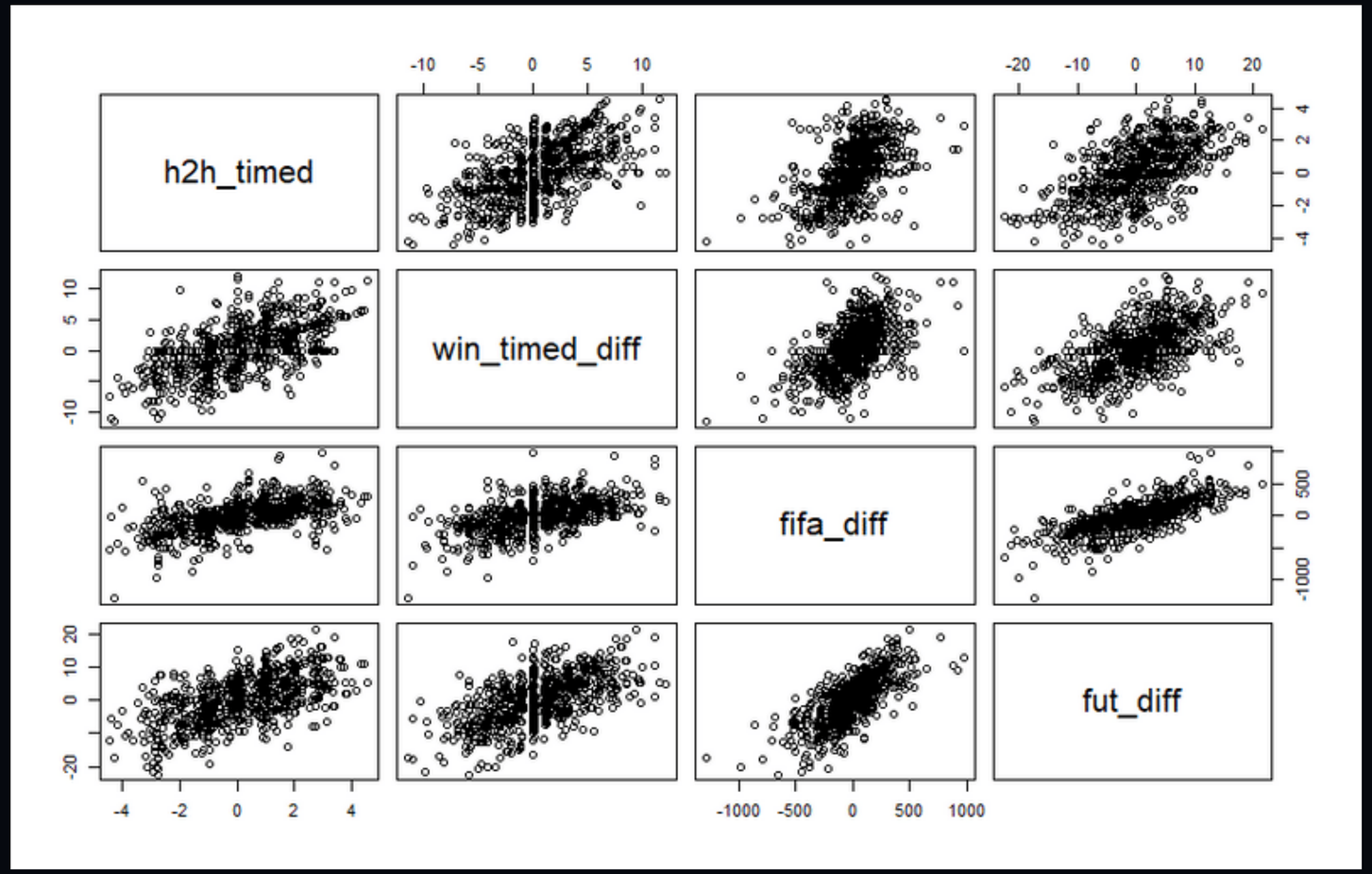
APPENDIX 6

Team	Quals
Mexico	1.000
Costa Rica	0.987
United States	0.968
Brazil	0.994
Argentina	0.993
Uruguay	0.965
Colombia	0.859
South Korea	0.792
Iran	0.685
Australia	0.850
Saudi Arabia	0.695
Senegal	0.651
Ivory Coast	0.639
Egypt	0.465
Ghana	0.462
DR Congo	0.446
Germany	0.998
Poland	0.998
England	0.994
France	0.914
Spain	0.897
Switzerland	0.865
Belgium	0.844
Italy	0.765
Serbia	0.749
Portugal	0.732
Croatia	0.541
Republic of Ireland	0.520
Sweden	0.381
Chile	0.696
Uzbekistan	0.677

2018						
Actual Qualification	Estimate From Last Year	Actually qualified?	Estimate Using FIFA Rankings	Actually qualified?	Our Model	Actually Qualified?
Russia	Brazil	TRUE	Russia	TRUE	Russia	TRUE
Brazil	Japan	TRUE	Australia	TRUE	Australia	TRUE
Iran	Australia	TRUE	Iran	TRUE	Iran	TRUE
Japan	Iran	TRUE	Saudi Arabia	TRUE	Saudi Arabia	TRUE
Mexico	South Korea	TRUE	South Korea	TRUE	South Korea	TRUE
Belgium	Netherlands	FALSE	Uzbekistan	FALSE	Uzbekistan	FALSE
South Korea	Italy	FALSE	Algeria	FALSE	DR Congo	FALSE
Saudi Arabia	Costa Rica	TRUE	Ghana	FALSE	Egypt	TRUE
Germany	United States	FALSE	Ivory Coast	FALSE	Ghana	FALSE
England	Argentina	TRUE	Senegal	TRUE	Ivory Coast	FALSE
Spain	Belgium	TRUE	Tunisia	TRUE	Tunisia	TRUE
Nigeria	Switzerland	TRUE	Mexico	TRUE	Mexico	TRUE
Costa Rica	Germany	TRUE	Costa Rica	TRUE	Costa Rica	TRUE
Poland	Colombia	TRUE	United States	FALSE	United States	FALSE
Egypt	Bosnia and Herzegovina	FALSE	Argentina	TRUE	Argentina	TRUE
Iceland	Russia	TRUE	Brazil	TRUE	Brazil	TRUE
Serbia	England	TRUE	Chile	FALSE	Chile	FALSE
Portugal	Spain	TRUE	Colombia	TRUE	Colombia	TRUE
France	Chile	FALSE	Uruguay	TRUE	Uruguay	TRUE
Uruguay	Ecuador	FALSE	Belgium	TRUE	Belgium	TRUE
Argentina	Honduras	FALSE	Croatia	TRUE	Croatia	TRUE
Colombia	Nigeria	TRUE	England	TRUE	England	TRUE
Panama	Ivory Coast	FALSE	France	TRUE	France	TRUE
Senegal	Cameroon	FALSE	Germany	TRUE	Germany	TRUE
Morocco	Ghana	FALSE	Iceland	TRUE	Italy	FALSE
Tunisia	Algeria	FALSE	Italy	FALSE	Poland	TRUE
Switzerland	Greece	FALSE	Netherlands	FALSE	Portugal	TRUE
Croatia	Croatia	TRUE	Poland	TRUE	Republic of Ireland	FALSE
Sweden	Portugal	TRUE	Portugal	TRUE	Serbia	TRUE
Denmark	France	TRUE	Spain	TRUE	Spain	TRUE
Australia	Mexico	TRUE	Switzerland	TRUE	Sweden	TRUE
Peru	Uruguay	TRUE	Wales	FALSE	Switzerland	TRUE
	20		23		24	

FIFA WORLD CUP Qatar2022

Pairs Plot Between Predictors



APPENDIX 7

FIFA WORLD CUP Qatar2022