UpGrad

FIFA WORLD CUP

RUSSIA 2018

CHALLENGE

CRYSTAL BALLERZ

PLAYERS

VIJAYANAND NARAYANAN

UPGRAD

SONALI CHHABRA

UPGRAD

VISHAL SHARMA

NON-UPGRAD

FIFA WORLD CUP 2018 CHALLENGE OVERVIEW

BACKGROUND: The most coveted tournament in football is the world cup that is conducted once every 4 years. 32 football playing nations will participate in the 2018 tournament held in Russia and one will emerge as the winner eventually.

OBJECTIVE: Carry out research, understand the teams, players, rankings and what it takes to win the tournament. Build a model predicting FIFA 2018 results.

- **Predict 4 semi-finalists**
- **Predict 2 finalists**
- **Predict winner of 2018 world cup**

RESULT: UFC PREDICTIONS

ROUND 16	QUARTER FINALS	SEMI FINALS		SEMI FINALS	QUARTER FINALS	ROUND 16
URUGUAY						SPAIN
	Portugal			Spain		
PORTUGAL		SF1 Portugal		SF3 Spain		RUSSIA
FRANCE						CROATIA
	France			Croatia		
ARGENTINA		F1 Brazil	BRAZIL CHAMPION	F2 Spain		DENMARK
BRAZIL						SWEDEN
	Brazil			Switzerland		
MEXICO		SF 2 Brazil		SF4 England		SWITZERLAND
BELGIUM						COLOMBIA
	Belgium			England		
JAPAN						ENGLAND

SEMI FINAL PREDICTIONS | FINAL PREDICTIONS | WINNER PREDICTIONS

DATASETS: FOUR KEY DATA SOURCES

1. **World Cup 2018 Matches – src: https://www.kaggle.com/agostontorok/soccer-world-cup-2018-winner/data This dataset consists match fixtures of all teams along with information on previous titles, previous finalists, semi finalists and their rankings**
2. **Results - src: https://www.kaggle.com/agostontorok/soccer-world-cup-2018-winner/data This dataset consists results of all football matches played between various countries from 1930 onwards**
3. **FIFA Rankings - src: https://www.kaggle.com/agostontorok/soccer-world-cup-2018-winner/data Data contains rankings of teams from 1993 onwards**
4. **Elo Ratings - src : https://www.eloratings.net/ This data contains Elo ratings of football teams based on the Elo rating system, developed by Dr. Arpad Elo. This rating has been chosen as it works well for a Zero Sum game such as football**

PROBLEM SOLVING APPROACH

Exploratory Data Analysis

Create Hypothesis for predictions

- **Analyze past win data**
- **Analyze trends across major countries**

Identify major data issues

Data Preparation

- **Data Cleaning**
- **Merge Data from all datasets**
- **Create derived metrics and dummy for categorical variables**

Model Building & Evaluation

- **Choose sample from dataset**
- **Choose model and train**
- **Feature Engineering**
- **Re-run model with selected features**
- **Fine tune performance of model**
- **Predict outcome of matches using test dataset**

EXPLORATORY DATA ANALYSIS

We ran Exploratory Data Analysis to understand historical data trends in order to build out a good hypothesis for our model building

- **Brazil, Germany, Argentina, Uruguay have been major past winners**
- **Brazil, Germany, Argentina, Uruguay and France have been in WC finals**
- **There are some strong teams like Belgium, Switzerland, Croatia that are strong teams, have not previously win but are strong contenders**
- **FIFA rankings are change but the last 3 years are good indicators of 'current team form'**

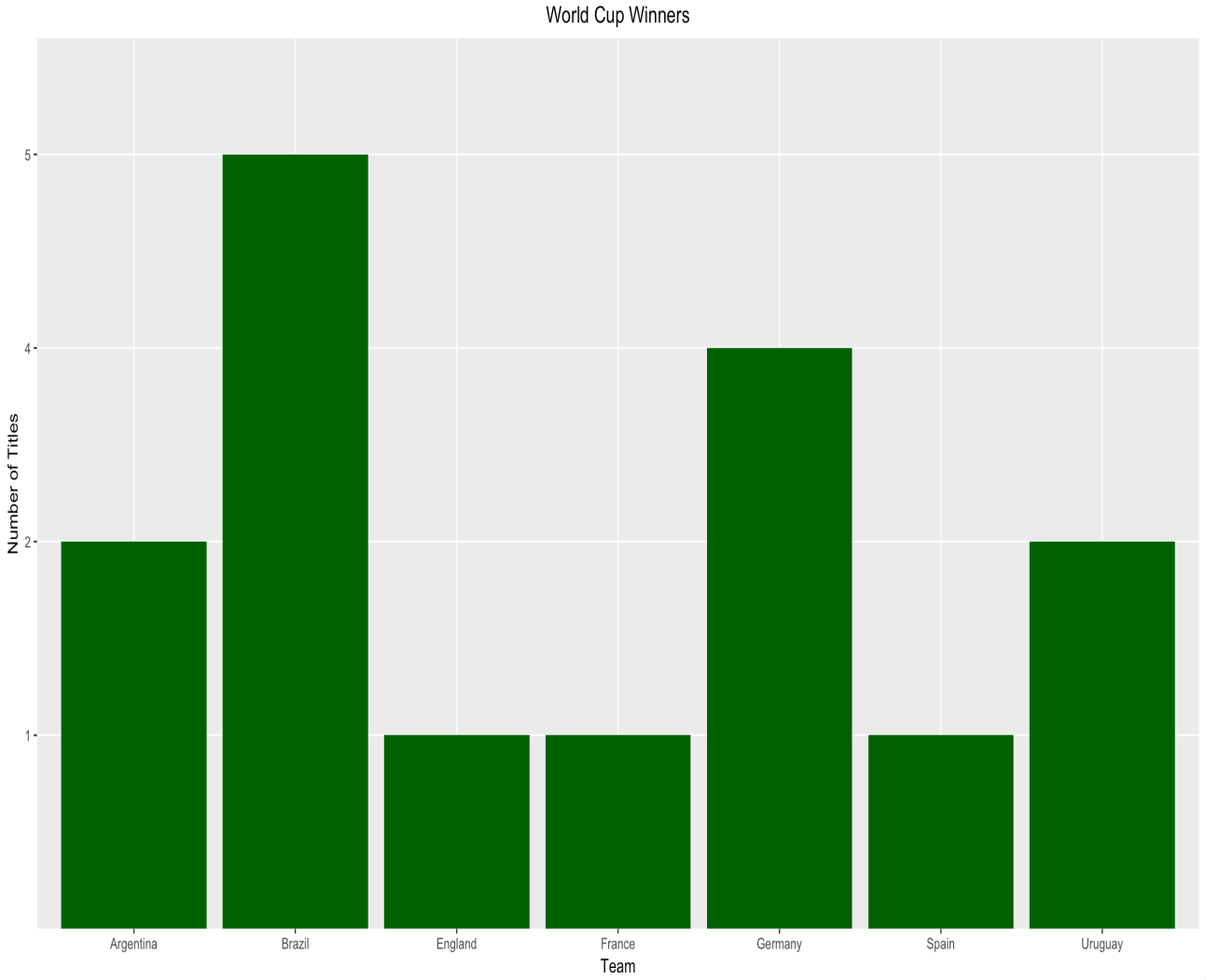
EDA: PAST WORLD CUP WINNERS

Past World Cup winners have been

Brazil – 5 times

Germany – 4 times

Argentina and Uruguay – 2 times



EDA: TEAMS IN PAST WC FINALS

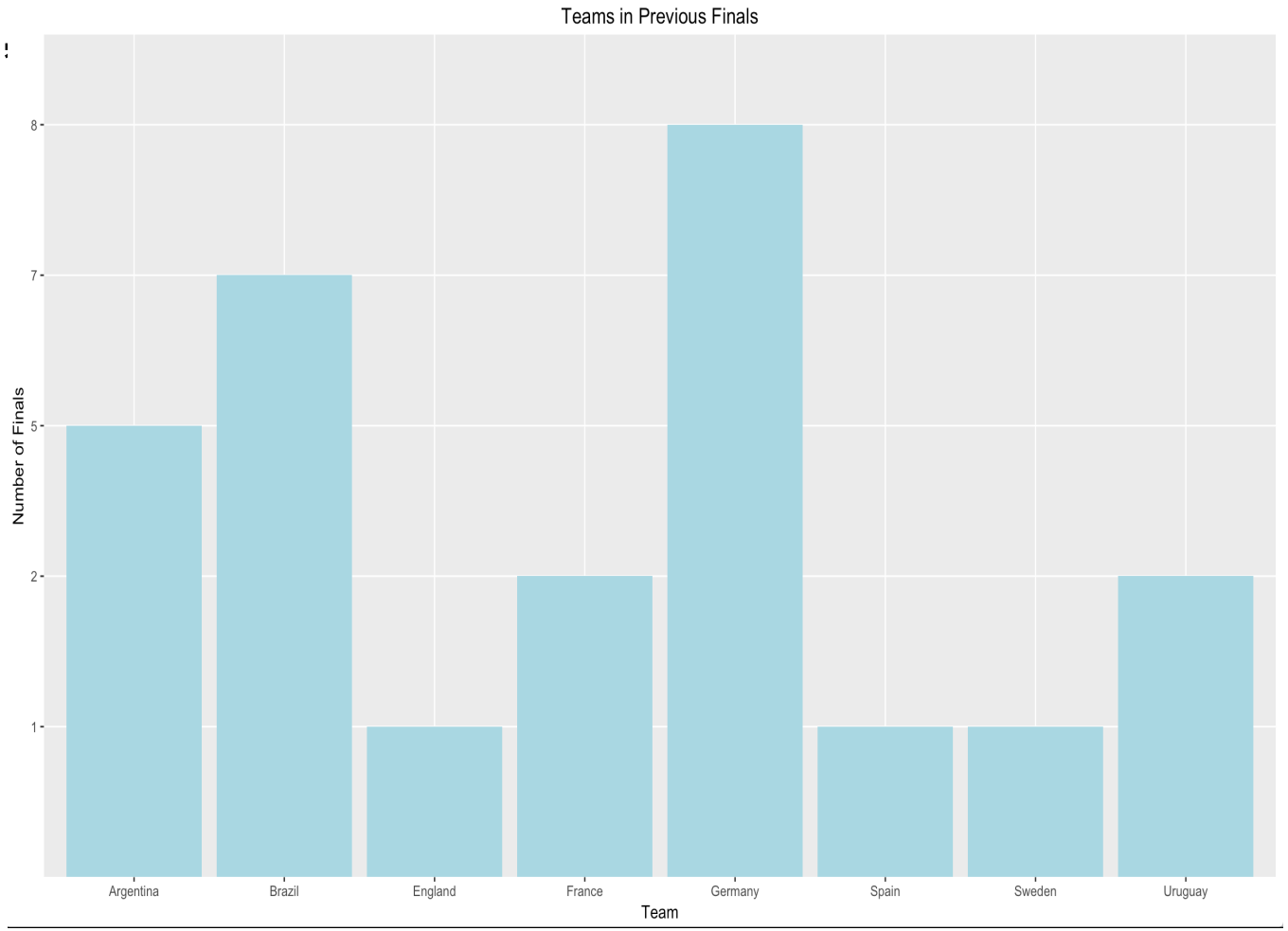
Teams that have been in previous finals:

Germany– 8 times

Brazil – 7 times

Argentina – 5 times

France and Uruguay – 2 times



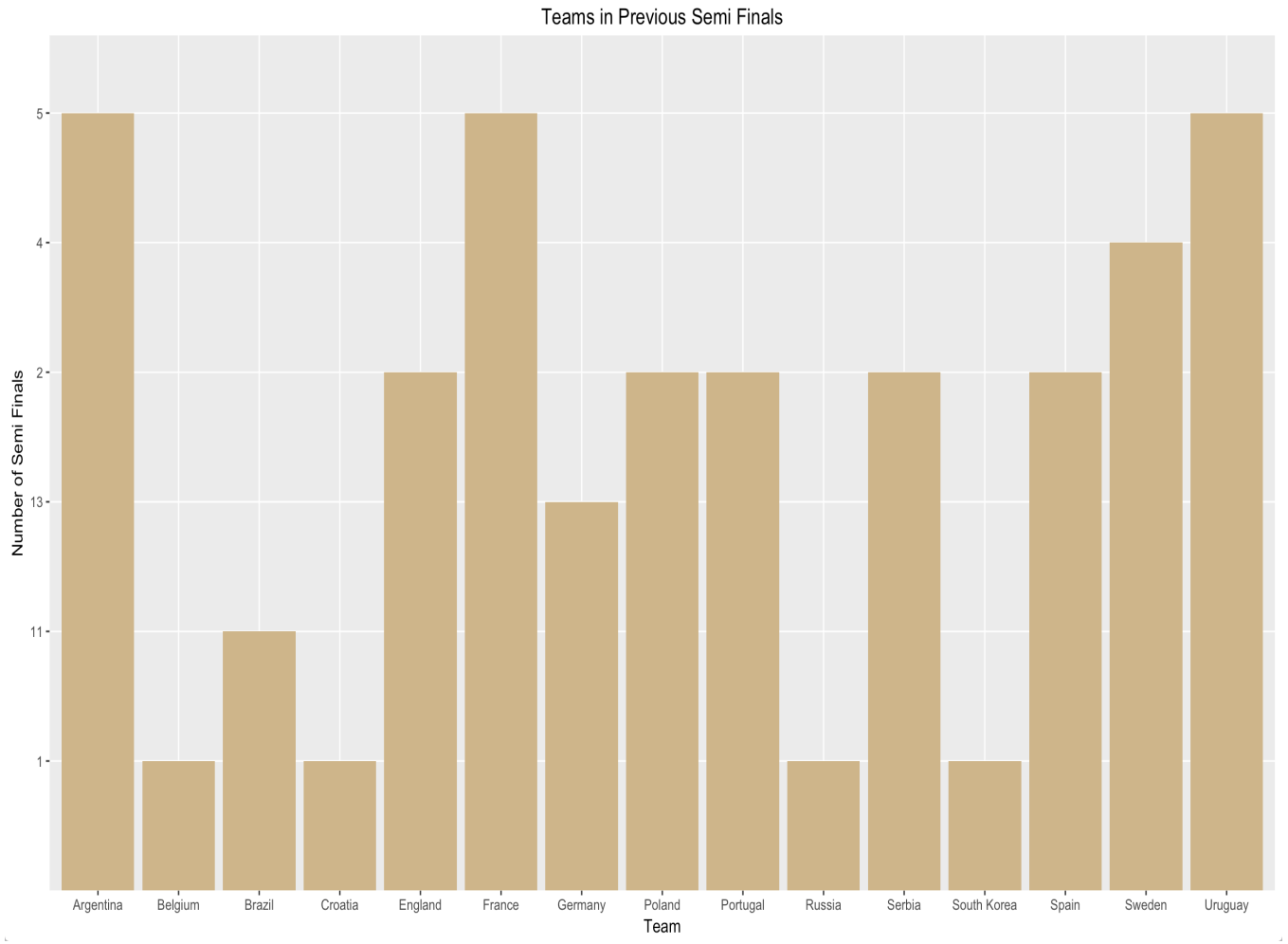
EDA: TEAMS IN PAST WC SEMI-FINALS

Teams that have been in previous semi finals,

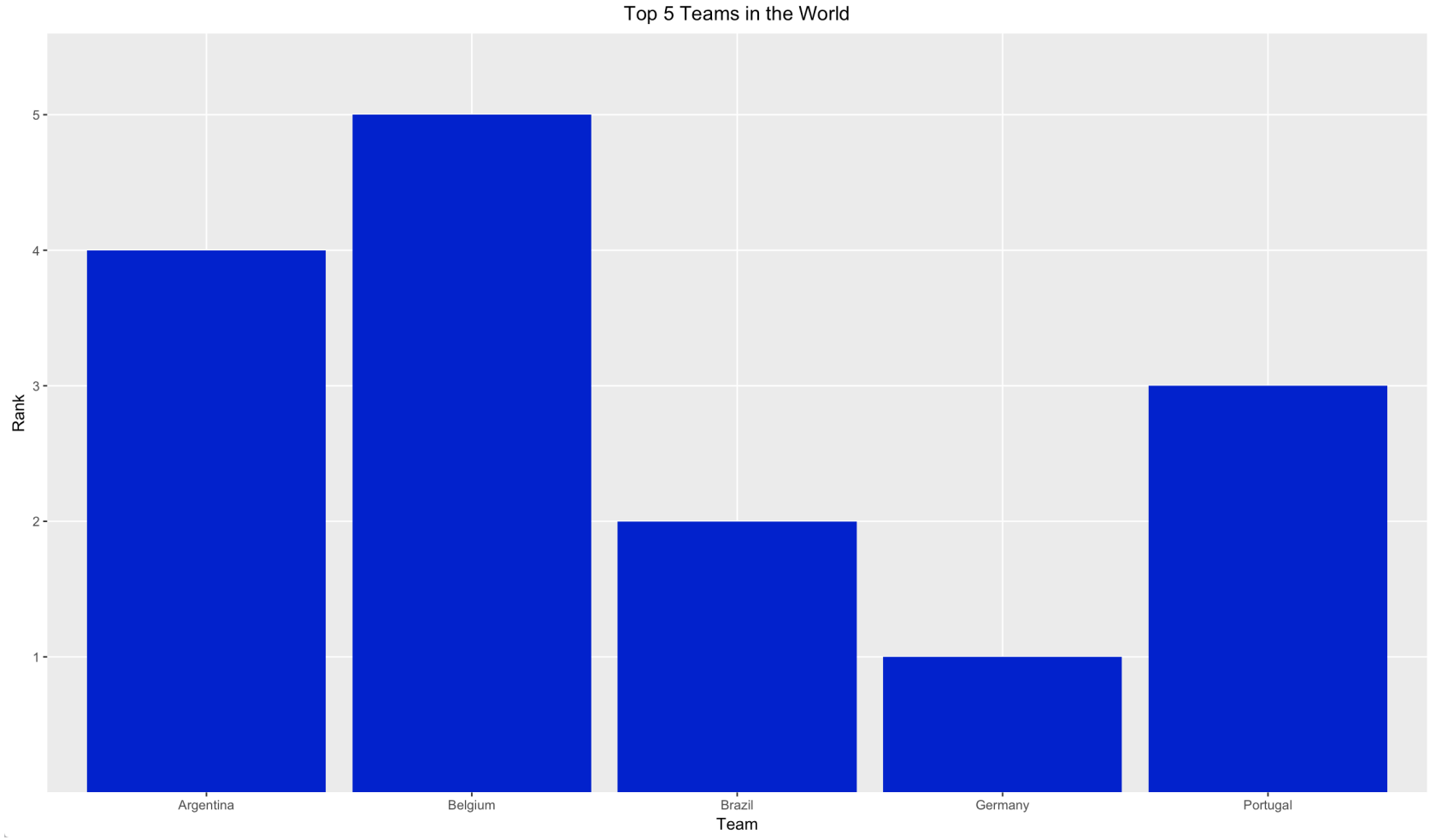
Germany – 13 times

Brazil – 11 times

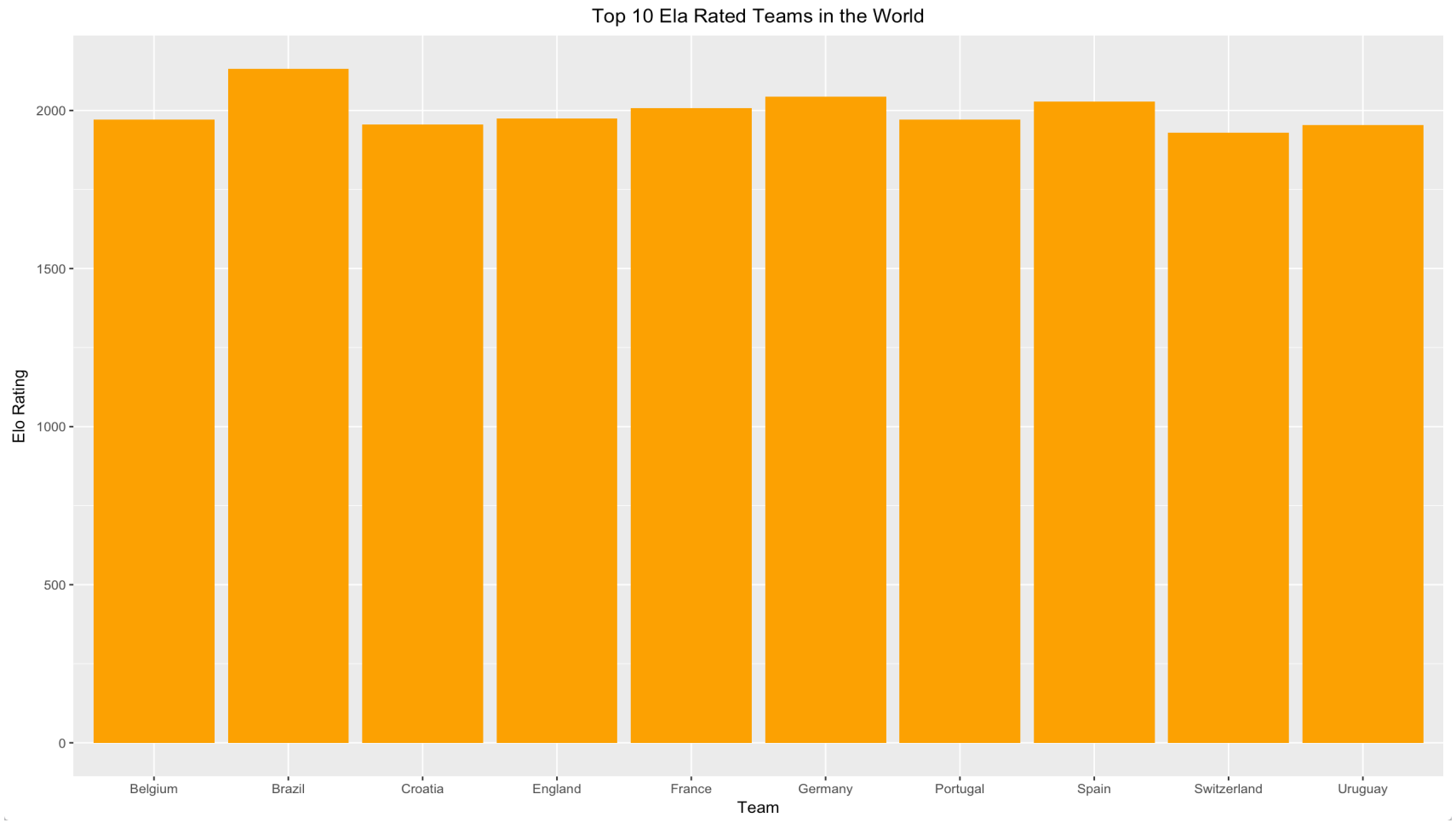
Argentina, France, Uruguay – 5 times



EDA: TOP 5 TEAMS IN THE WORLD BY FIFA RANKING



EDA: TOP 10 ELO RATED TEAMS IN THE WORLD



DATA PREPARATION: CLEANING

Following steps were taken to clean data,

- **Consistent team names. For example Portugal instead Porugal**
- **Check for duplicate observations**
- **Handle missing values**
- **Replace incorrect values**
- **Re-format date**

DATA PREPARATION: MERGING

Considerations when combining data

- **Select records from dataset from year 1993 onwards. This is because FIFA rankings are available only from 1993**
- **If a country has multiple rankings for a year in the dataset then select the most recent one**
- **If Elo rating is missing for a team then replace with a default value of 1300 (recommended default value for a football team)**
- **Remove columns such as City, Country, Match Year as they are irrelevant for the prediction**

MODEL BUILDING: TARGET VARIABLE

Target (dependent) variable for the prediction is:

MATCH OUTCOME

- **Target variable for a group match can have 3 values i.e 1-home team win, 0-home team loss or 0.5-draw**
- **Target variable for knockout stage matches can only have 1-home team win or 0-home team loss**

MODEL BUILDING: MODEL SELECTION

Three Models were tried out to make predictions,

1. **Logistic Regression**
2. **SVM and**
3. **Random Forest**

From the above, Random Forest was chosen to be the final model because of better accuracy, sensitivity and specificity values over the other models.

MODEL BUILDING: RANDOM FOREST

Exploratory Data Analysis → Data Preparation → Model Building & Evaluation

```
randomForestModel <- randomForest(outcome ~ ., train, ntree = 30000, mtry = 5,
nodesize = 0.01 * nrow(train))
```

Confusion Matrix

```
          Reference
Prediction  0 0.5  1
       0   20   8 13
       0.5  8  14 14
       1    3   8 54
```

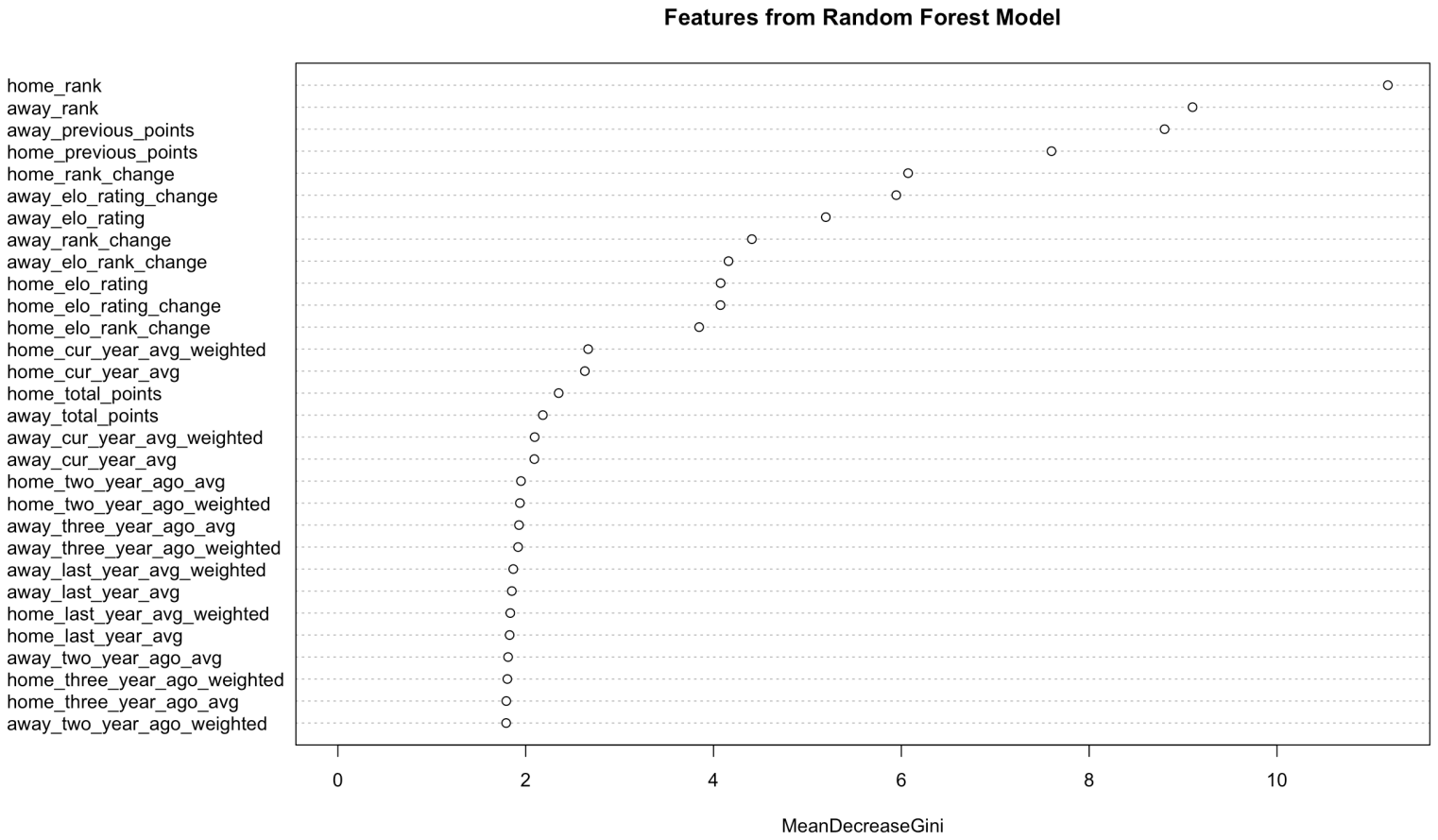
Accuracy: 61.97%,

```
Statistics by Class:

                     Class: 0 Class: 0.5 Class: 1
Sensitivity            0.6452    0.46667   0.6667
Specificity            0.8108    0.80357   0.8197
Pos Pred Value         0.4878    0.38889   0.8308
Neg Pred Value         0.8911    0.84906   0.6494
Prevalence             0.2183    0.21127   0.5704
Detection Rate         0.1408    0.09859   0.3803
Detection Prevalence   0.2887    0.25352   0.4577
Balanced Accuracy      0.7280    0.63512   0.7432
```

MODEL BUILDING: FEATURE ENGINEERING

Exploratory Data Analysis → Data Preparation → Model Building & Evaluation



MODEL BUILDING: IMPORTANT FEATURES

outcome ~ home_rank + away_rank + home_elo_rating + away_elo_rating + home_rank_change + away_rank_change + home_elo_rank_change + away_elo_rank_change + home_elo_rating_change + away_elo_rating_change + home_cur_year_avg + away_cur_year_avg + home_cur_year_avg_weighted + away_cur_year_avg_weighted + home_two_year_ago_avg + away_two_year_ago_avg + home_two_year_ago_weighted + away_two_year_ago_weighted + home_three_year_ago_avg + away_three_year_ago_avg + home_three_year_ago_weighted + away_three_year_ago_weighted + home_total_points + away_total_points + home_previous_points + away_previous_points

MODEL BUILDING: RANDOM FOREST WITH SELECTED FEATURES

UpGrad

Exploratory Data Analysis → Data Preparation → Model Building & Evaluation

```
randomForestModel_1 <- randomForest(formula_1, train, ntree = 30000, mtry = 2,
nodesize = 0.01 * nrow(train))
```

Confusion Matrix

```
          Reference
Prediction  0 0.5  1
       0   22   7 12
       0.5  6  14 16
       1    3   6 56
```

Accuracy: 64.79%

,

```
Statistics by Class:

                     Class: 0 Class: 0.5 Class: 1
Sensitivity            0.7097    0.51852   0.6667
Specificity            0.8288    0.80870   0.8448
Pos Pred Value         0.5366    0.38889   0.8615
Neg Pred Value         0.9109    0.87736   0.6364
Prevalence             0.2183    0.19014   0.5915
Detection Rate         0.1549    0.09859   0.3944
Detection Prevalence   0.2887    0.25352   0.4577
Balanced Accuracy      0.7693    0.66361   0.7557
```

MODEL OUTPUT: GROUP STAGE PREDICTIONS

A sample of the group stage predictions is shown

44 out out 48 group matches were predicted correctly (shown in green)

4 out of the 48 matches were predicted incorrectly (shown in red)

home_team	away_team	actual_result	pred_outcome
Portugal	Spain	0.5	0.5
Portugal	Morocco	1	1
Iran	Spain	0	0
Iran	Portugal	0.5	0.5
Spain	Morocco	0.5	0.5
France	Australia	1	1
Peru	Denmark	0	0
Denmark	Australia	0.5	0.5
France	Peru	1	1
Denmark	France	0.5	0.5
Australia	Peru	0	0
Argentina	Iceland	0.5	1
Croatia	Nigeria	1	1
Argentina	Croatia	0	0
Nigeria	Iceland	1	1
Nigeria	Argentina	0	0
Iceland	Croatia	0	0
Costa Rica	Serbia	0	0.5
Brazil	Switzerland	0.5	0.5
Brazil	Costa Rica	1	1
Serbia	Switzerland	0	0
Serbia	Brazil	0	0
Switzerland	Costa Rica	0.5	0.5
Germany	Mexico	0	1
Sweden	South Korea	1	1
South Korea	Mexico	0	0
Germany	Sweden	1	1
Mexico	Sweden	0	0
South Korea	Germany	1	0
Belgium	Panama	1	1
Tunisia	England	0	0
Belgium	Tunisia	1	1
England	Panama	1	1

RESULT: UFC PREDICTIONS

ROUND 16	QUARTER FINALS	SEMI FINALS		SEMI FINALS	QUARTER FINALS	ROUND 16
URUGUAY						SPAIN
	Portugal			Spain		
PORTUGAL		SF1 Portugal		SF3 Spain		RUSSIA
FRANCE						CROATIA
	France			Croatia		
ARGENTINA		F1 Brazil	BRAZIL CHAMPION	F2 Spain		DENMARK
BRAZIL						SWEDEN
	Brazil			Switzerland		
MEXICO		SF 2 Brazil		SF4 England		SWITZERLAND
BELGIUM						COLOMBIA
	Belgium			England		
JAPAN						ENGLAND

SEMI FINAL PREDICTIONS | FINAL PREDICTIONS | WINNER PREDICTIONS