FIFA Challenge

Introduction

This is notebook for Upgrad FIFA prediction challenge. we have used 4 datasets, 6 features & weighted average method to make naive predictions for round 16, quarter-finals, semi-finals & finals. visualization was done in tableau & remianing eda, data preprocessing & score calculation was done in R.

Importing Datasets

Data Sources:

1. https://www.kaggle.com/tadhgfitzgerald/fifa-international-soccer-mens-ranking-1993now (fifa_ranking.csv)
2. https://www.kaggle.com/ahmedelnaggar/fifa-worldcup-2018-dataset (World Cup 2018 Dataset.csv)
3. https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017 (results.csv)
4. https://github.com/neaorin/PredictTheWorldCup/tree/master/input (matches.csv)

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(scales)

results <- read.csv("results.csv") #Match details of different teams
fifa_ranking <- read.csv("fifa_ranking.csv") #Yearwise fifa rankings of various teams
matches <- read.csv("matches.csv")
fifa2018 <- read.csv("World Cup 2018 Dataset.csv") #ranks & other details of teams participating in FIF
```

Feature Extraction

We will extract key features from above dataframes first to do some exploratoty analysis & later include complex features for analysis

```
fifa_stats <- subset(fifa2018,select=c(Team,Current..FIFA.rank,Previous..titles,Previous..finals,Previou
```

Exploratory Analysis

Top Ten Champions competing in FIFA 2018

```
#head(sort(fifa_stats$Current..FIFA.rank), n = 10)
x = arrange(fifa_stats,Current..FIFA.rank)
x = x[1:10,]
head(x,10)
```

```
##           Team Current..FIFA.rank Previous..titles Previous..finals
## 1      Germany                  1                4                8
## 2       Brazil                  2                5                7
## 3      Porugal                  3                0                0
## 4    Argentina                  4                2                5
## 5      Belgium                  5                0                0
## 6        Spain                  6                1                1
## 7       Poland                  7                0                0
## 8  Switzerland                  8                0                0
## 9       France                  9                1                2
## 10        Peru                 11                0                0
##    Previous..semifinals
## 1                    13
## 2                    11
## 3                     2
## 4                     5
## 5                     1
## 6                     2
## 7                     2
## 8                     0
## 9                     5
## 10                    0
```

Visualization: FIFA 2018 Tableau Dashboard https://public.tableau.com/profile/mohit5191#!/vizhome/FIFA2018ExploratotyAnalysis/Dashboard1

Lets check average ranking based on team ranking from 1993 to 2018

```
AverageRank <- fifa_ranking %>% group_by(country_full) %>% summarise(mean(rank)) %>% setNames(c('Country
AverageRank <- as.data.frame(AverageRank)
head(AverageRank,10)
```

```
##        Country AverageRank
## 1       Brazil    3.171329
## 2      Germany    5.104895
## 3        Spain    5.321678
## 4    Argentina    5.454545
## 5        Italy    8.353147
## 6  Netherlands    8.888112
## 7       France    8.958042
## 8      England   10.653846
## 9     Portugal   11.346154
## 10      Mexico   14.751748
```

Checking Average total points of teams

```
AverageTotalPoints <- fifa_ranking %>% group_by(country_full) %>% summarise(mean(total_points)) %>% setN
AverageTotalPoints <- as.data.frame(AverageTotalPoints)
head(AverageTotalPoints,10)
```

```
##       Country AverageTotalPoints
## 1     Germany           421.5051
## 2      Serbia           397.2937
## 3   Argentina           395.2343
## 4       Spain           389.8979
```

```
## 5  Montenegro           368.0464
## 6      Brazil           363.5058
## 7    Portugal           349.5588
## 8    Colombia           339.0694
## 9     Belgium           334.8879
## 10    Uruguay           324.1228
```

Analyzing matches among various teams

```
levels(matches$CupName)
```

```
## [1] "Confederation competition team final"
## [2] "FIFA competition team final"
## [3] "FIFA competition team qualification"
## [4] "Friendly"
```

Lets filter only matches from "FIFA competition team final"

```
fifa_matches <- subset(matches,CupName="FIFA competition team final")
```

Finding total goals scored by different teams

```
TotalGoalsT1 <- fifa_matches %>% group_by(team1Text) %>% summarise(sum(team1Score)) %>% setNames(c('Cou

TotalGoalsT2 <- fifa_matches %>% group_by(team2Text) %>% summarise(sum(team2Score)) %>% setNames(c('Cou
```

Finding lead maintained by various teams during past FIFA teams

```
fifa_matches$team1Lead <- fifa_matches$team1Score - fifa_matches$team2Score
fifa_matches$team2Lead <- fifa_matches$team2Score - fifa_matches$team1Score

fifa_matches_team1 <- subset(fifa_matches,select=c(team1Text,team1Lead))
fifa_matches_team2 <- subset(fifa_matches,select=c(team2Text,team2Lead))

fifa_matches_team1_avg <- fifa_matches_team1 %>% group_by(team1Text) %>% summarise(mean(team1Lead, na.r

fifa_matches_team2_avg <- fifa_matches_team2 %>% group_by(team2Text) %>% summarise(mean(team2Lead, na.r

fifa_matches_teams_avg <- merge(fifa_matches_team1_avg, fifa_matches_team2_avg, by = "Country")
fifa_matches_teams_avg$AverageLead <- fifa_matches_teams_avg$AverageLead.x + fifa_matches_teams_avg$Ave
fifa_matches_teams_avg <- subset(fifa_matches_teams_avg,select=c(Country,AverageLead))
```

Lets combine these features & analyse further

```
fifa <- data_frame()

fifa <- merge(fifa_stats,AverageRank,by.x = "Team",by.y = "Country")
fifa <- merge(fifa,AverageTotalPoints,by.x = "Team",by.y = "Country")

fifa <- merge(fifa,fifa_matches_teams_avg,by.x = "Team",by.y = "Country")
```

So final selected features for prediction are Previous Titles, Previous Finals, Previous Semifinals, Average Total Points, Average Lead & Average Rank. however these features need further prerocessing.

Further Preprocessing

```
fifa <- fifa[,-c(2)]
fifa$InverseAvgRank <- 1/fifa$AverageRank
fifa <- fifa[,-c(5)]
```

Scaling Data

This step is neccessary as we will be using a weighted average for predictions, here we are scaling all 6 features on scle of 1 to 10

```
fifa_subset <- fifa[,c(2:7)]
fifa_subset <- data.frame(lapply(fifa_subset,function(x) rescale(as.numeric(x), to=c(1,10))))

summary(fifa_subset)
```

```
##  Previous..titles Previous..finals Previous..semifinals AverageTotalPoints
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000       Min.   : 1.000
##  1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.000       1st Qu.: 3.001
##  Median : 1.000   Median : 1.000   Median : 1.000       Median : 4.405
##  Mean   : 2.067   Mean   : 2.125   Mean   : 2.385       Mean   : 5.129
##  3rd Qu.: 1.900   3rd Qu.: 2.125   3rd Qu.: 2.385       3rd Qu.: 6.922
##  Max.   :10.000   Max.   :10.000   Max.   :10.000       Max.   :10.000
##   AverageLead     InverseAvgRank
##  Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 5.640   1st Qu.: 1.373
##  Median : 6.222   Median : 1.612
##  Mean   : 6.221   Mean   : 2.574
##  3rd Qu.: 7.185   3rd Qu.: 2.400
##  Max.   :10.000   Max.   :10.000
```

Calculating weighted winning likelihood

Weightages given to features are:

Feature	Weightage
Previous Titles	5%
Previous Finals	10%
Previous Semifinals	10%
Average Total Points	25%
Average Lead	25%
Average Rank	25%

```
fifa <- cbind(fifa[,c(1)],fifa_subset)
names(fifa)[names(fifa) == "fifa[, c(1)]"] <- "Team"

WinningWeightage = c()
for(i in 1:nrow(fifa)){
WinningScore <- sum(fifa[i,2]*.05,fifa[i,3]*.1,fifa[i,4]*.1,fifa[i,5]*.25,fifa[i,6]*.25,fifa[i,7]*.25)
WinningWeightage <- c(WinningWeightage, WinningScore)
}

fifa$WinChance <- WinningWeightage
```

Calculating individual match winners using WinChance, predicted winners are mentioned in comment after each comparision

Current Team Positions

Round 16:

```
which(fifa[fifa$Team=="France",]$WinChance > fifa[fifa$Team=="Argentina",]$WinChance) #Argentina
```

```
## integer(0)
```

```
#which(fifa[fifa$Team=="Uruguay",]$WinChance > fifa[fifa$Team=="Porugal",]$WinChance) Porugal

which(fifa[fifa$Team=="Brazil",]$WinChance > fifa[fifa$Team=="Mexico",]$WinChance) #Brazil
```

```
## [1] 1
```

```
which(fifa[fifa$Team=="Belgium",]$WinChance > fifa[fifa$Team=="Japan",]$WinChance) #Belgium
```

```
## [1] 1
```

```
which(fifa[fifa$Team=="Spain",]$WinChance > fifa[fifa$Team=="Russia",]$WinChance) #Spain
```

```
## [1] 1
```

```
which(fifa[fifa$Team=="Croatia",]$WinChance > fifa[fifa$Team=="Denmark",]$WinChance) #Croatia
```

```
## [1] 1
```

```
which(fifa[fifa$Team=="Sweden",]$WinChance > fifa[fifa$Team=="Switzerland",]$WinChance) #Sweden
```

```
## [1] 1
```

```
#which(fifa[fifa$Team=="Columbia",]$WinChance > fifa[fifa$Team=="England",]$WinChance) #England
```

Quarter-Finals:

```
which(fifa[fifa$Team=="Porugal",]$WinChance > fifa[fifa$Team=="Argentina",]$WinChance) #Argentina
```

```
## integer(0)
```

```
which(fifa[fifa$Team=="Brazil",]$WinChance > fifa[fifa$Team=="Belgium",]$WinChance) #Brazil
```

```
## [1] 1
```

```
which(fifa[fifa$Team=="Spain",]$WinChance > fifa[fifa$Team=="Croatia",]$WinChance) #Spain
```

```
## [1] 1
```

```
which(fifa[fifa$Team=="Sweden",]$WinChance > fifa[fifa$Team=="England",]$WinChance) #England
```

```
## integer(0)
```

Semi-Finals:

```
which(fifa[fifa$Team=="Brazil",]$WinChance > fifa[fifa$Team=="Argentina",]$WinChance) #Brazil
```

```
## [1] 1
```

```
which(fifa[fifa$Team=="Spain",]$WinChance > fifa[fifa$Team=="England",]$WinChance) #Spain
```

```
## [1] 1
```

Finals:

```
which(fifa[fifa$Team=="Spain",]$WinChance > fifa[fifa$Team=="Brazil",]$WinChance) #Brazil
```

```
## integer(0)
```

So Brzail should win FIFA 2018 as per our naive approach.

File with final features used to make predictions is saved as final_dataset.csv in output folder

```
#write.csv(fifa,"final_dataset.csv")
```