

STAT 107 World Cup Prediction

Team 23 - Jigar Patel, Henry Yost, Davinia Muthalaly, Refugio Zepeda

2025-11-07

Abstract

For our project, the main question we are trying to answer is, using previous years data set, are we able to pull accurate, predictable factors/variables which can guide us in being able to predict a future World Cup winner. What we are more specifically trying to look for, is next years World Cup winner by building a model around that statistics we have. From our preliminary work, we have found some strong variables, that lead to which teams advance to the next round and which don't. We have also looked into some distributions for advancing teams and non-advancing teams to see different relationships in the data set for multiple previous World Cups.

Introduction

Our purpose for this analysis is to make an efficient model that can get close to, if not, precisely predict the World Cup winner. We want to be able to achieve this goal by using only previous years data sets and the information that comes with it. This analysis can benefit many who are interested in the world of soccer and have a passion for the game. We will try to build a model in which we train it to take in the variables from our data set we want to use, and be able to use the information to make an informed hypothesis of the next World Cup winner.

Data

Our data set was created by Ph.D developer who was interested in the World Cup. This was all accessible through a github repository that we found in our search for reliable data. We have many different variables for this dataset from team ID's to wins/losses/draws, as well as goals_for, goals_against, games played, and a couple more informative and performance based variables.

- Necessary Libraries

```
source("00_libraries.R")
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.1.0
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr    1.3.1
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()   masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

- Reading in Datasets
- Pre Processing (Data Cleaning/Processing)

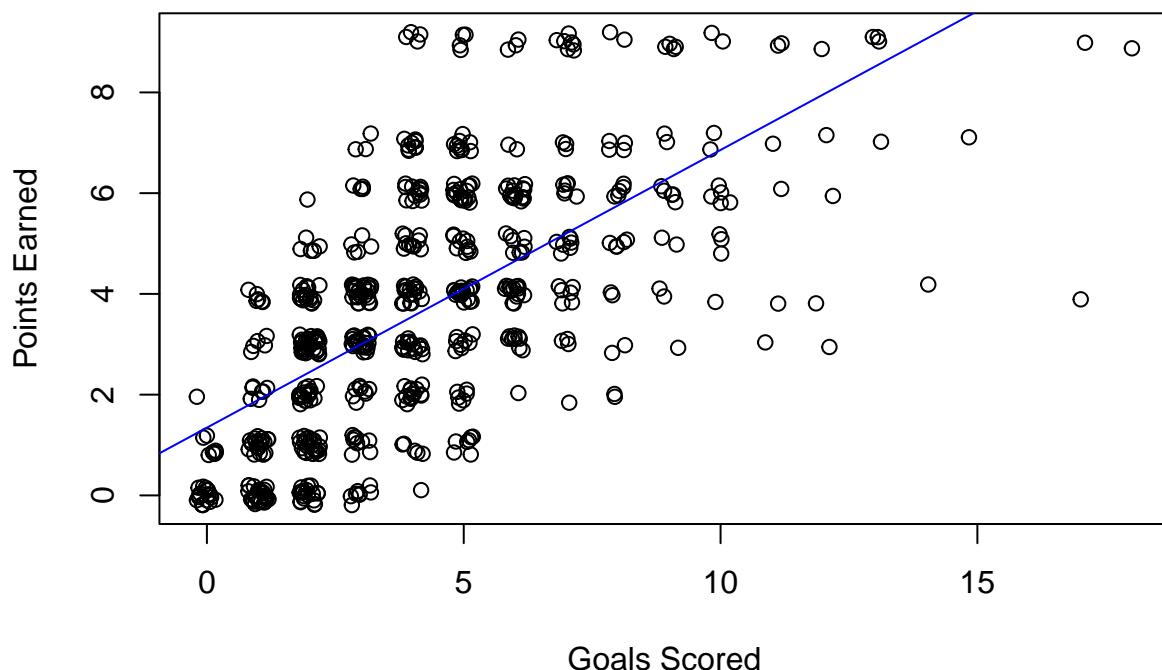
```
source("10_DataCleaning.R")
```

Visulaization

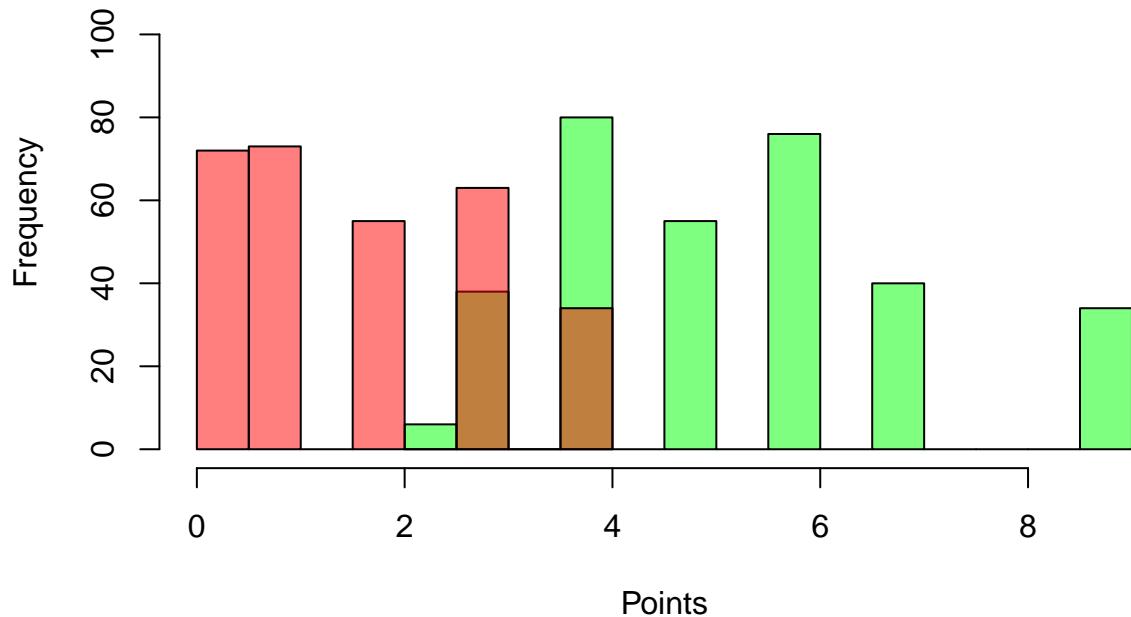
- Preliminary EDA

```
source("11_PreliminaryEDA.R")
```

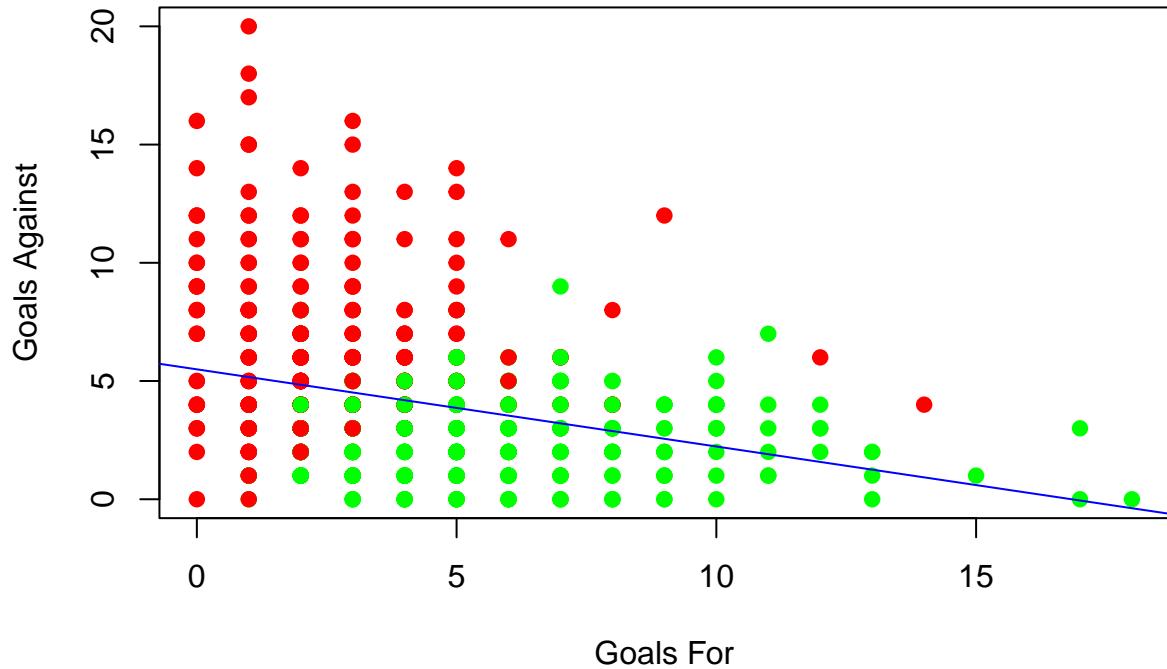
Relationship between goals scored and points



Points Distribution: Advanced vs Non-Advanced



Goals For vs Goals Against



```

## Start:  AIC=288.6
## advanced ~ played + wins + draws + losses + goals_for + goal_difference +
##           goals_against + points
##
##
## Step:  AIC=288.6
## advanced ~ played + wins + draws + losses + goals_for + goal_difference +
##           points
##
##
## Step:  AIC=288.6
## advanced ~ played + wins + draws + goals_for + goal_difference +
##           points
##
##          Df Deviance    AIC
## - goals_for     1   56.671 286.80
## - points        1   56.741 287.58
## - goal_difference 1   56.780 288.01
## <none>          56.653 288.60
## - played         1   57.484 295.72
## - wins          1   62.945 352.54
## - draws          1   64.034 363.27
##
## Step:  AIC=286.8
## advanced ~ played + wins + draws + goal_difference + points
##

```

```

##               Df Deviance   AIC
## - points      1  56.758 285.77
## - goal_difference  1  56.787 286.09
## <none>          56.671 286.80
## - played       1  57.485 293.73
## - wins          1  62.959 350.68
## - draws         1  64.189 362.79
##
## Step:  AIC=285.77
## advanced ~ played + wins + draws + goal_difference
##
##               Df Deviance   AIC
## - goal_difference  1  56.876 285.06
## <none>          56.758 285.77
## - played         1  57.505 291.95
## - draws          1  70.307 417.78
## - wins          1  84.856 535.51
##
## Step:  AIC=285.06
## advanced ~ played + wins + draws
##
##               Df Deviance   AIC
## <none>          56.876 285.06
## - played        1  57.520 290.12
## - draws         1  75.330 458.97
## - wins          1 154.660 909.29
## Start:  AIC=367.67
## advanced ~ goals_for + goal_difference + goals_against + points
##
##
## Step:  AIC=367.67
## advanced ~ goals_for + goal_difference + points
##
##               Df Deviance   AIC
## - goals_for      1  64.985 366.50
## <none>          64.899 367.67
## - goal_difference  1  66.592 381.79
## - points         1  85.544 538.57
##
## Step:  AIC=366.5
## advanced ~ goal_difference + points
##
##               Df Deviance   AIC
## <none>          64.985 366.50
## - goal_difference  1  66.912 382.79
## - points         1  85.545 536.58

```

Analysis

Relationship between goals scored and points

- **Plot Analysis (Relationship between Goals Scored and Points)** This graph plots the relationship between goals scored and points throughout the tournament(s), and there exists an expectation; teams that

score more goals will tend to earn more points. Thus, given this plot gives a positive linear relationship the plot behaves as an initial sanity check of the dataset. Looking into the graph, we can see that most teams score less than 10 goals scored throughout the tournament, suggesting the majority of the teams have a strong defense or the games are low-scoring (a hypothesis worth exploring further). Additionally, a large cluster of teams is around the 4 points and less than 5 goals scored, this shows us that majority of teams do not advance far into the tournament. Conversely, teams outside these ranges like correspond to teams progressing to the later stages of the tournament, which can be an invaluable trend that can help towards predicting future games.

Histogram - Points Distribution: Advanced vs. Not Advanced

- Histogram Analysis (Advanced vs. Not Advanced) This histogram shows us the point distribution for all the teams in the World Cup. The red and green differentiate teams that advanced to the next round, which is green, and the teams that did not advance to the next round, which is in red. As we can see from the histogram, the max points a team that didn't advance was 4 and the minimum a team got that did advance was about. Logically this would make sense, as better teams would show a stronger defense in matches, therefore the teams that didn't advance would produce lower points in those matches. And vice versa for points scored. A better team would also have a stronger offense, therefore their points in the histogram would show as much higher than the teams that didn't advance. This is a good insight to look at because we can base our future prediction off of higher scoring teams, as it shows some relation to advancing teams.

Scatterplot - Goals For vs. Goals Against

- Scatter Plot Analysis (Goals For vs. Goals Against) Something to note for this scatter plot to begin with is the bottom-left represents a low scoring game (defensively, low scoring teams). The top-left shows a high scoring game. The bottom right corner shows a strong defensive and offensive game which is the ideal game for advancing teams. And the top right corner would represent the opposite, a weak game, both defensively and offensively. From the scatter plot we see the green, or advancing teams, are congregating in the bottom right corner of the plot while the red, or not advancing teams, are more clustered from very bottom left to top right corner. This is another good insight to note for advancing teams will normally have higher goals_for and lower goals_against for them to advance. The goals against is something we can also look at for future predictions we want to make about the world cup.

Logistic Regression (Best prediction factors)

- Logistic Regression Analysis (Model) Using logistic regression on the group standings performance variables, we found three variables to be best predictors out of all of the variables. Played games, draws, and wins were the three variables that created the lowest model AIC from all the variables. Using backwards step wise selection, we take out any variables that has a higher AIC than the model AIC because once removed, the model AIC goes lower. And from that we reached a model with the three variables mentioned earlier as great predictors for whether a team advances or not.

- Logistic Regression Analysis (Model 2) We made another model, however only using 4 of the variables to look a more performance based model. From this we see that the last model shows the variables goal_difference and points. These two variables made for the lowest AIC score and the best predictors out of the 4 variables used in the second model. We see that the second models AIC score (366.50) is higher than the first models AIC score (285.06), however they capture different results. The first model provides variables that show direct outcomes of a game like wins and draws. The second model provides variables that use the performance of the team in a game, such as points scored and the goal_difference between the teams for that game.