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 year's World Cup winner by building a model around the data 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

Note: We currently have an additional dataset (Sofascore) that contains more data about individual games and team ELO. However, this dataset was originally structured for use in Python, and requires us to move it into R.

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

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v purrr     1.1.0
## v forcats   1.0.1     v stringr   1.5.2
## v ggplot2   4.0.0     v tibble    3.3.0
## 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 Dataset + Preprocessing (Data Cleaning/Processing)

```
# function called that cleans group_standings.csv by stripping non-numeric,
# and uses RegEx to convert to numeric (if applicable)
df_clean = clean_group_standings("data/group_standings.csv")
cat("Dataset size: ", nrow(df_clean), "rows and", ncol(df_clean), "columns\n")
```

```
## Dataset size:  626 rows and 13 columns
```

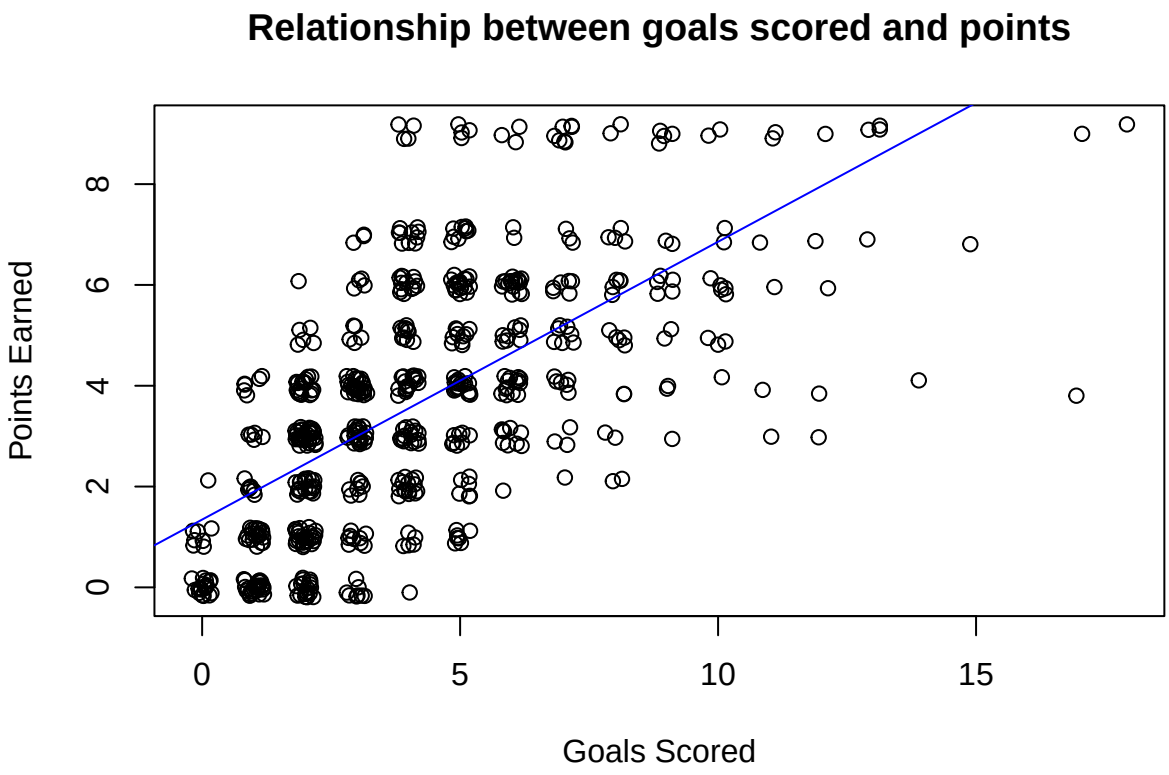
```
head(df_clean, 10)
```

```
## # A tibble: 10 x 13
##    tournament_id team_name  team_code played  wins draws losses goals_for
##    <chr>         <chr>      <chr>      <dbl> <dbl> <dbl>  <dbl>     <dbl>
##  1 WC-1930       Argentina  ARG            3     3     0      0        10
##  2 WC-1930       Chile      CHL            3     2     0      1         5
##  3 WC-1930       France     FRA            3     1     0      2         4
##  4 WC-1930       Mexico     MEX            3     0     0      3         4
##  5 WC-1930       Yugoslavia YUG            2     2     0      0         6
##  6 WC-1930       Brazil     BRA            2     1     0      1         5
##  7 WC-1930       Bolivia    BOL            2     0     0      2         0
##  8 WC-1930       Uruguay    URY            2     2     0      0         5
##  9 WC-1930       Romania    ROU            2     1     0      1         3
## 10 WC-1930       Peru       PER            2     0     0      2         1
## # i 5 more variables: goals_against <dbl>, goal_difference <dbl>, points <dbl>,
## #   advanced <dbl>, year <chr>
```

This chunk is to show the Dataset size (626 x 13) and also give a sneak-peek into the dataset, so that its easier to understand what sort of data is being used in the document.

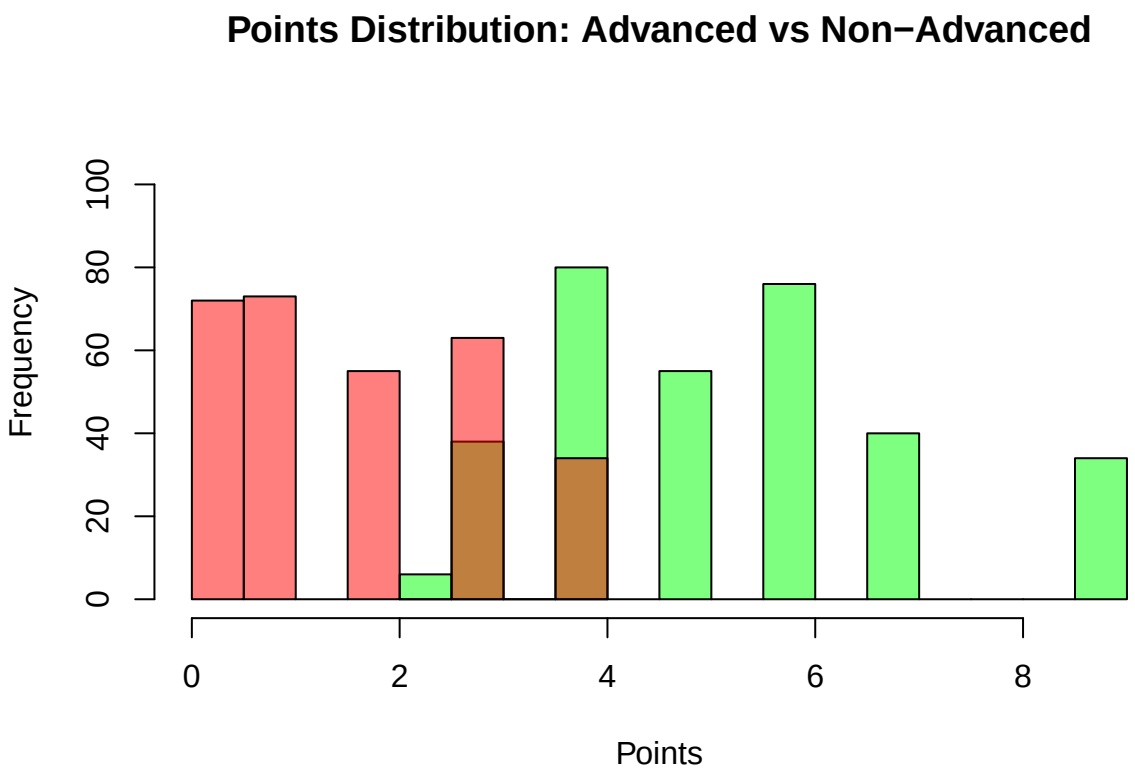
Visualization

Preliminary Exploratory Data Analysis (EDA)

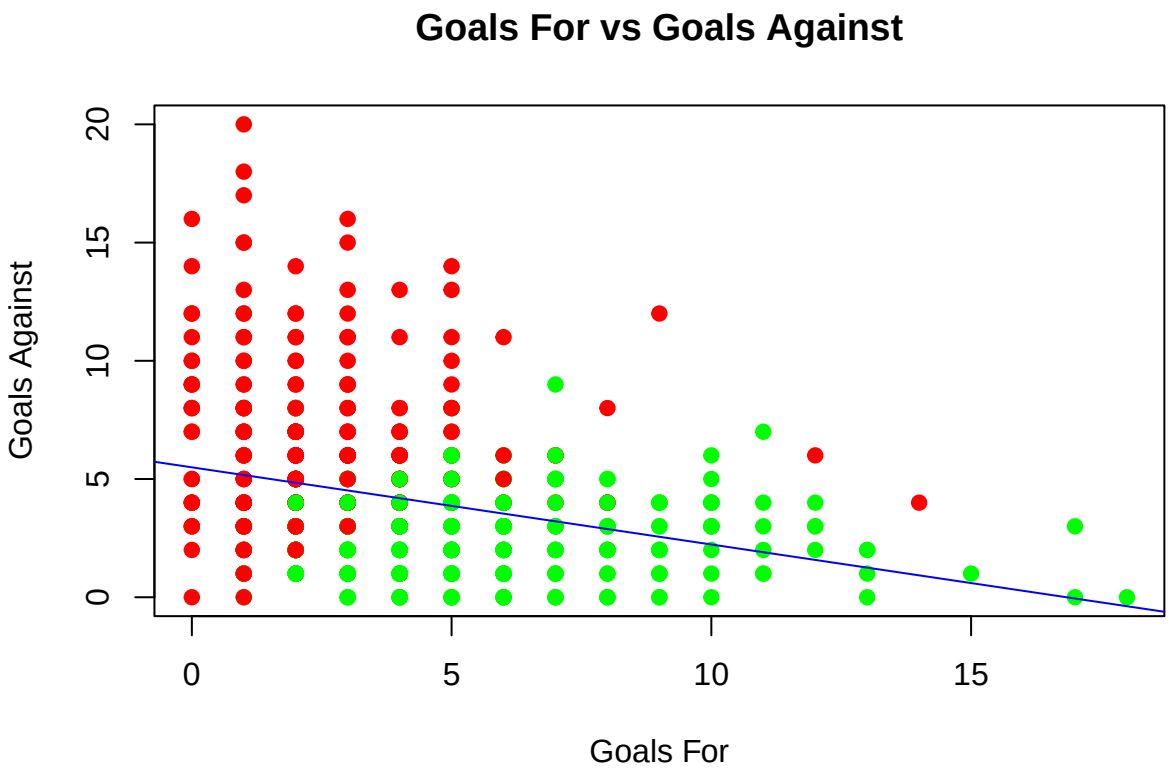



```
##
## Call:
## lm(formula = points ~ goals_for, data = data)
##
## Coefficients:
## (Intercept)    goals_for
##      1.3447       0.5512
```

Scatter Plot Analysis - Relationship between Goals Scored and Points Earned: 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 Analysis - Points Distribution: 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 Analysis - Goals For vs. Goals Against: In this scatter plot, the bottom-left represents low-scoring games (defensive 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.

```
# Some factor levels produced warnings due to perfect separation.
# results still indicate the main predictive variables.

models <- run_logistic_regression(df_clean)
summary(models$reduced_model1)
summary(models$reduced_model2)
```

Logistic Regression 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.

We made another model, however only using 4 of the variables to look a more performance based model.

From this we see that the second 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.

Analysis

For the goal of predicting the next World Cup winner, we plan to implement a more sophisticated model that is able to capture and take advantage of the non-linear relationships in the dataset between multiple variables. Right now, we are leaning towards an XGBoost (Extreme Gradient Boosting) model, because they are effective for tabular data where variables interact in complex ways (such as the soccer dataset). It is also known for being highly predictive and flexible with proper tuning, which hopefully will help us get a more accurate and consistent result.

Additionally, we may explore feature engineering, such as calculating team specific variables (momentum, win ratios, etc...) which could provide extra valuable predictive power. Additionally, Cross validation and proper evaluation metrics will be used to validate and confirm the accuracy of our model performance.

Currently, our next steps are to get the data from Sofascore integrated into R, and that will give us significantly more information to work with. Additionally, we want to create binary labels for each tournament winner for training.