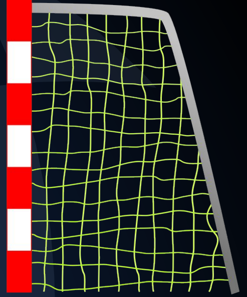
THE ULTIMATE SOCCER PREDICTION MODEL

Ben Wieland, Victor Vassallo, Arushi Shah, Reed Baumgardner



INTRODUCTION

FIFA Rankings are neither a good representation nor predictor of team success, we feel that we can do better.

Coca-Cola | FIFA MEN'S WORLD RANKING = 🗑

Table of Contents

1 Research Question

2 Data Collection

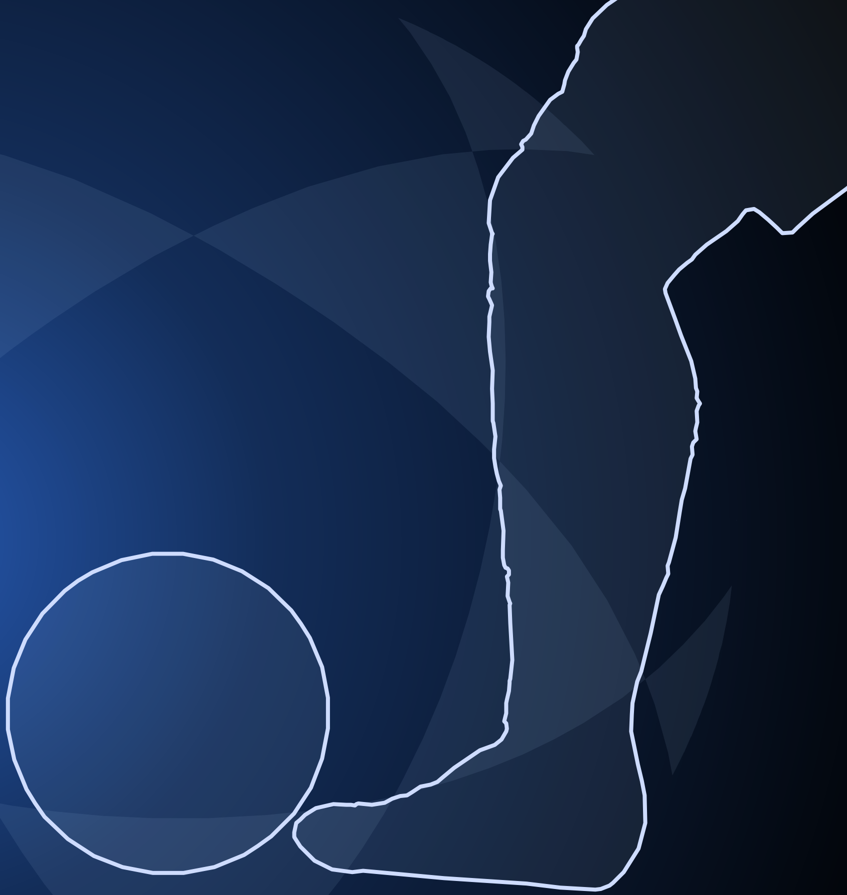
3 Analysis Plan

4 Justification

I

RESEARCH QUESTION

Can we create a predictive international soccer rankings system?



2

Data Collection

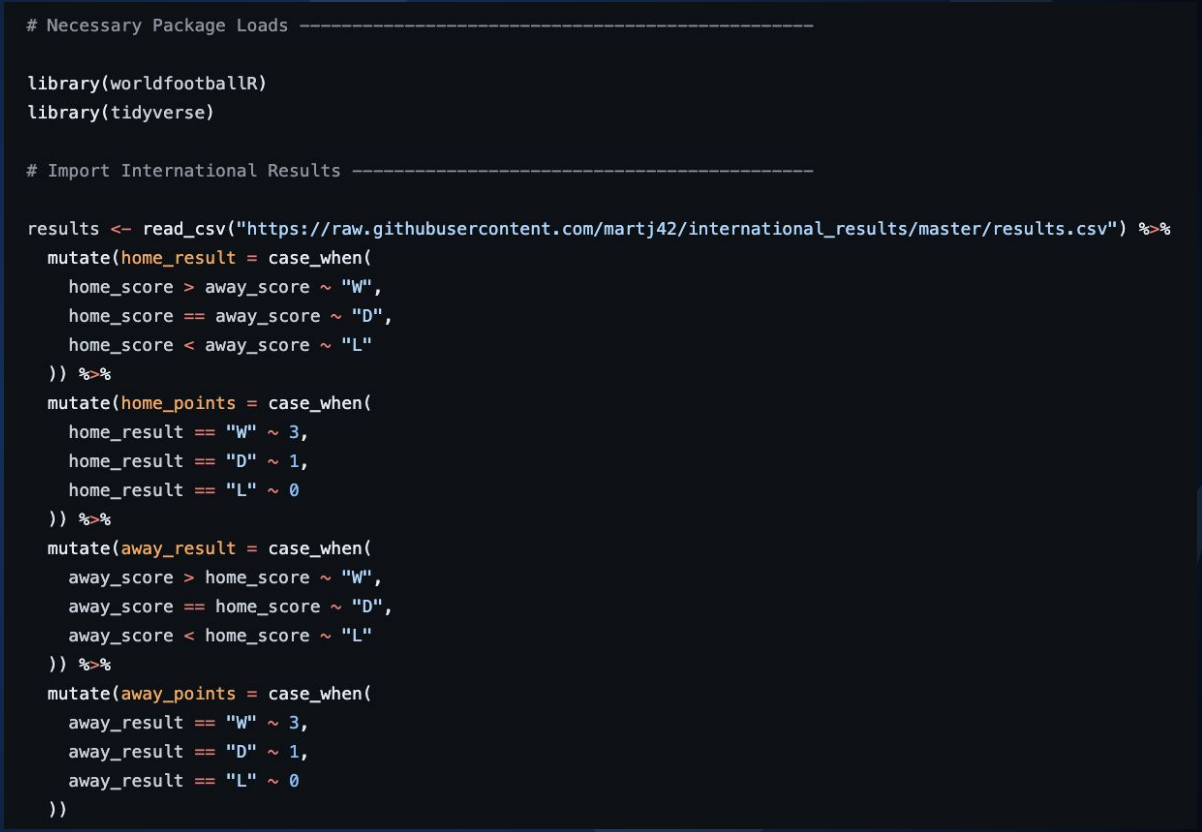
Sources

- Data collected from variety of sources, but primarily through the worldfootballR package provided by Github user JaseZiv
  - Package provided built-in functions to scrape data from a number of popular soccer statistics sites including:
    - Fbref.com
    - Transfermarkt.com
    - Understat.com
  - Package also provided functions to load pre-scraped match data
- To collect data specifically on international match results we used a Kaggle dataset that recorded every international soccer match between the years of 1872 and 2023
  - Date
  - Teams
  - Score
  - Competition
  - Location

Variables

- Our primary goal throughout the data collection stage was to cumulate statistics that would most effectively contribute to offensive, defensive, and overall ratings for each national team
    - Home vs away
        - Neutral location
    - Tournament
    - Goals for vs goals against
    - Expected goals
    - Age
    - Minutes played
    - Yellow cards
    - Red cards
    - Match date

Loading Data



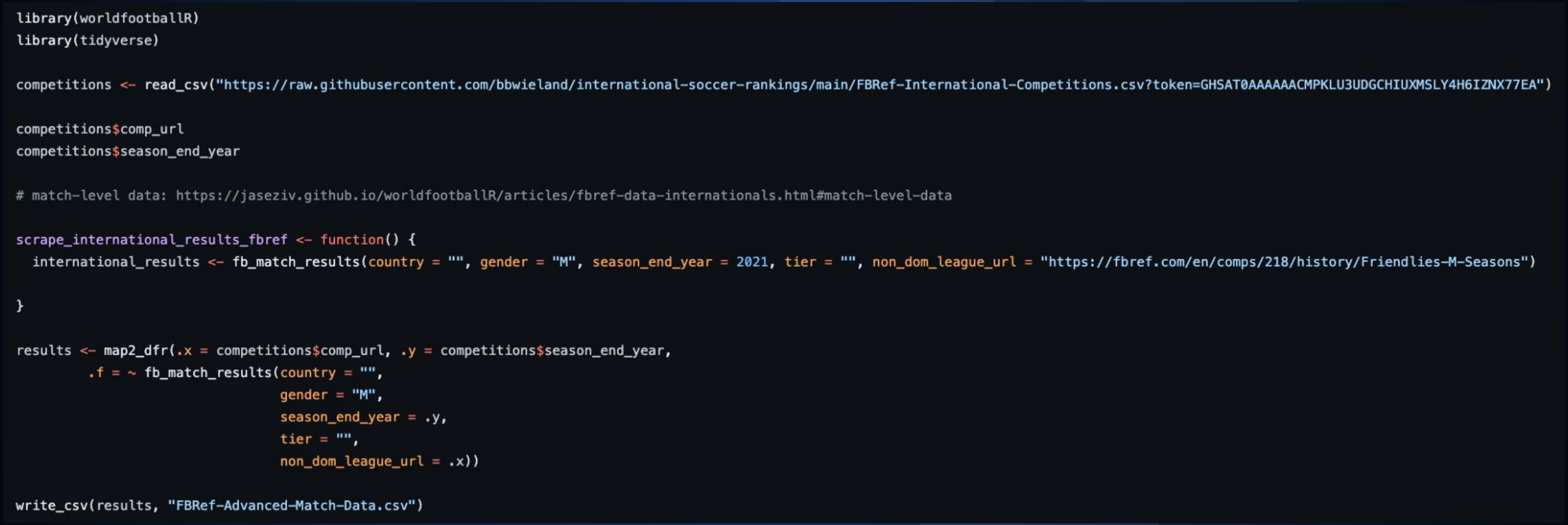
```
# Necessary Package Loads --------------------------------------------------

library(worldfootballR)
library(tidyverse)

# Import International Results ---------------------------------------------

results <- read_csv("https://raw.githubusercontent.com/martj42/international_results/master/results.csv") %>%
  mutate(home_result = case_when(
    home_score > away_score ~ "W",
    home_score == away_score ~ "D",
    home_score < away_score ~ "L"
  )) %>%
  mutate(home_points = case_when(
    home_result == "W" ~ 3,
    home_result == "D" ~ 1,
    home_result == "L" ~ 0
  )) %>%
  mutate(away_result = case_when(
    away_score > home_score ~ "W",
    away_score == home_score ~ "D",
    away_score < home_score ~ "L"
  )) %>%
  mutate(away_points = case_when(
    away_result == "W" ~ 3,
    away_result == "D" ~ 1,
    away_result == "L" ~ 0
  ))
```

Scraping Data



```r
library(worldfootballR)
library(tidyverse)

competitions <- read_csv("https://raw.githubusercontent.com/bbwieland/international-soccer-rankings/main/FBRef-International-Competitions.csv?token=GHSAT0AAAAAACMPKLU3UDGCHIUXMSLY4H6IZNX77EA")

competitions$comp_url
competitions$season_end_year

# match-level data: https://jaseziv.github.io/worldfootballR/articles/fbref-data-internationals.html#match-level-data

scrape_international_results_fbref <- function() {
  international_results <- fb_match_results(country = "", gender = "M", season_end_year = 2021, tier = "", non_dom_league_url = "https://fbref.com/en/comps/218/history/Friendlies-M-Seasons")

}

results <- map2_dfr(.x = competitions$comp_url, .y = competitions$season_end_year,
         .f = ~ fb_match_results(country = "",
                                 gender = "M",
                                 season_end_year = .y,
                                 tier = "",
                                 non_dom_league_url = .x))

write_csv(results, "FBRef-Advanced-Match-Data.csv")
```

Challenges

- Varying degrees of data collected between international competitions and countries
- Detailed statistics only appeared in the last thirty years
- Necessity of data to be cleaned and tweaked before merging and analysis
- Difficult to account for the differing levels of competition between every international tournament
- Avoiding similar developed country biases held by FIFA

3

Analysis Plan

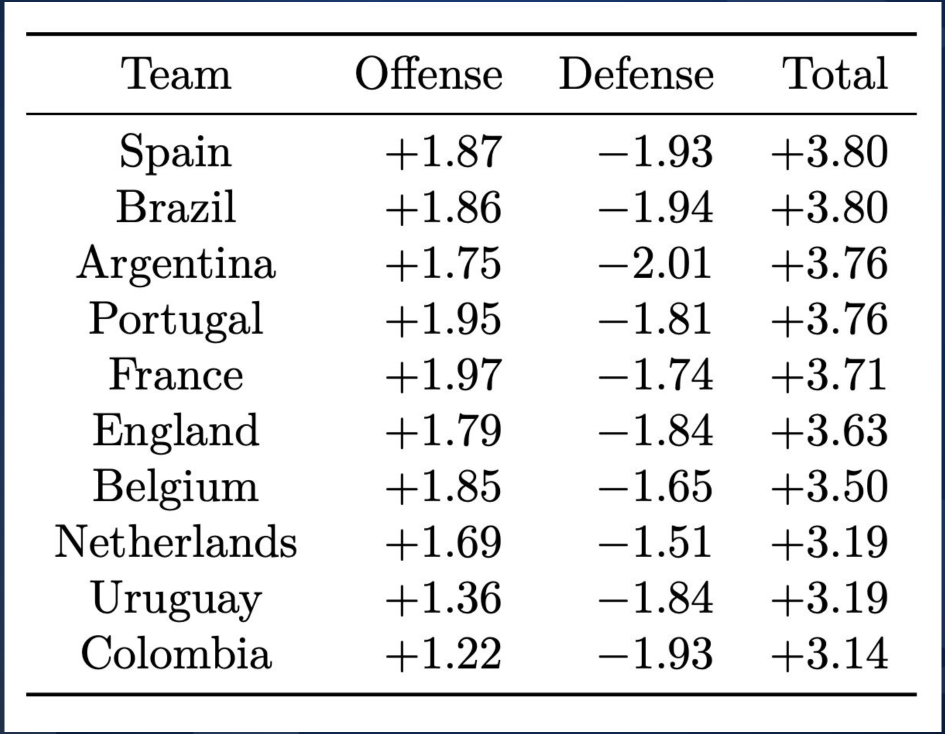
Two Key Analysis Questions:

1. Given our observed data, what is the best way to determine team strengths in a predictive manner?
   a. Output: Separate values for offensive & defensive team abilities.
2. Given those team strengths, how do we predict individual match results?
   a. Output: If we know who the two teams in a match are, as well as which team is playing at home, then we should be able to calculate the probabilities of each team winning, losing, or drawing the match.

Calculating Team Strengths

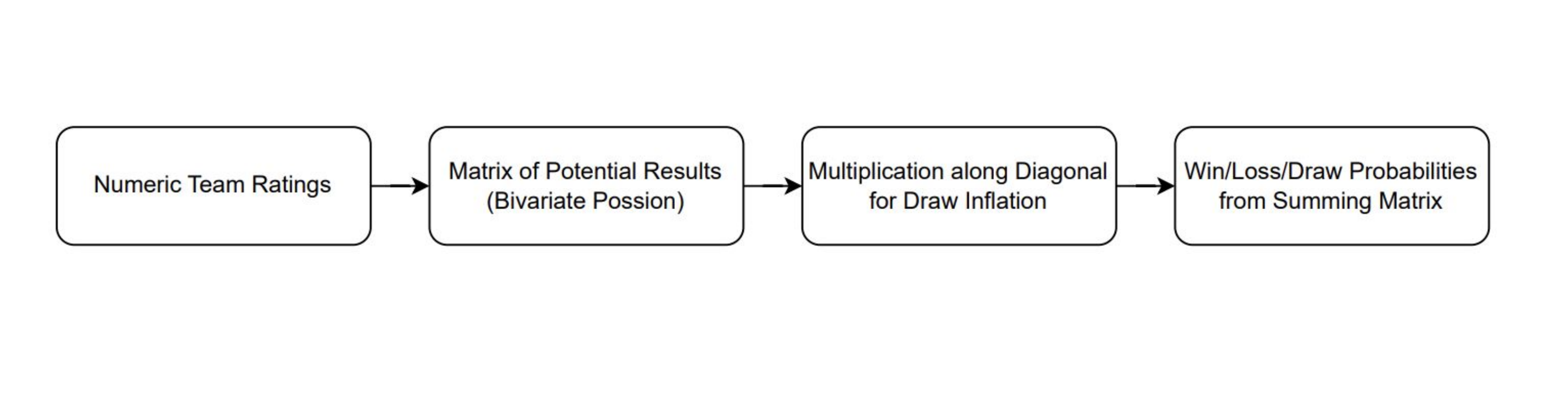
- Using previous match results:
  - Mixed-effects modeling: obtain most likely team strengths given final scores
  - Exponential decay (half-life) to weight recent results more heavily in the model
  - Advantage: straightforward & yields reasonable results
  - Disadvantage: struggles with teams in "bubbles" and to project mismatches
- Using information about players on each roster
  - Transfermarkt player valuations; the team as sum of its player values
  - Allows us to better project matches where teams don't field typical roster
- Data availability varies by match type (more complete for World Cup than friendlies)
  - Requires a variety of model types to handle different data availability
- Ideal final product: a blend of our various team- and player-level models

Example: Simple Mixed-Effects Time Decay Model



Team	Offense	Defense	Total
Spain	+1.87	−1.93	+3.80
Brazil	+1.86	−1.94	+3.80
Argentina	+1.75	−2.01	+3.76
Portugal	+1.95	−1.81	+3.76
France	+1.97	−1.74	+3.71
England	+1.79	−1.84	+3.63
Belgium	+1.85	−1.65	+3.50
Netherlands	+1.69	−1.51	+3.19
Uruguay	+1.36	−1.84	+3.19
Colombia	+1.22	−1.93	+3.14

From Team Strengths to Match Probabilities



Numeric Team Ratings

Matrix of Potential Results (Bivariate Possion)

Multiplication along Diagonal for Draw Inflation

Win/Loss/Draw Probabilities from Summing Matrix

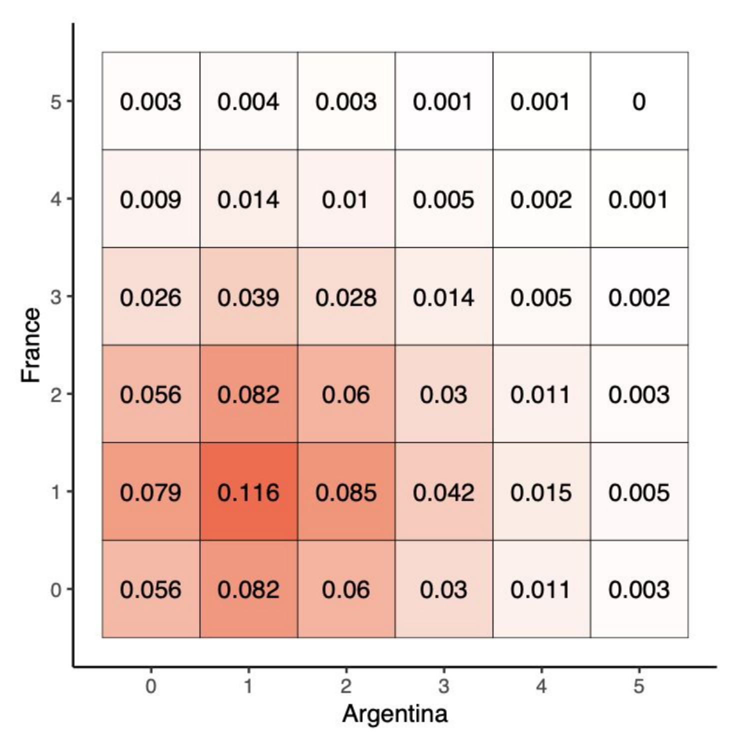
Example: 2022 World Cup Final

Using the team effects for Argentina and France, we calculated the following values:

ARGENTINA: 1.47 expected goals
FRANCE: 1.41 expected goals

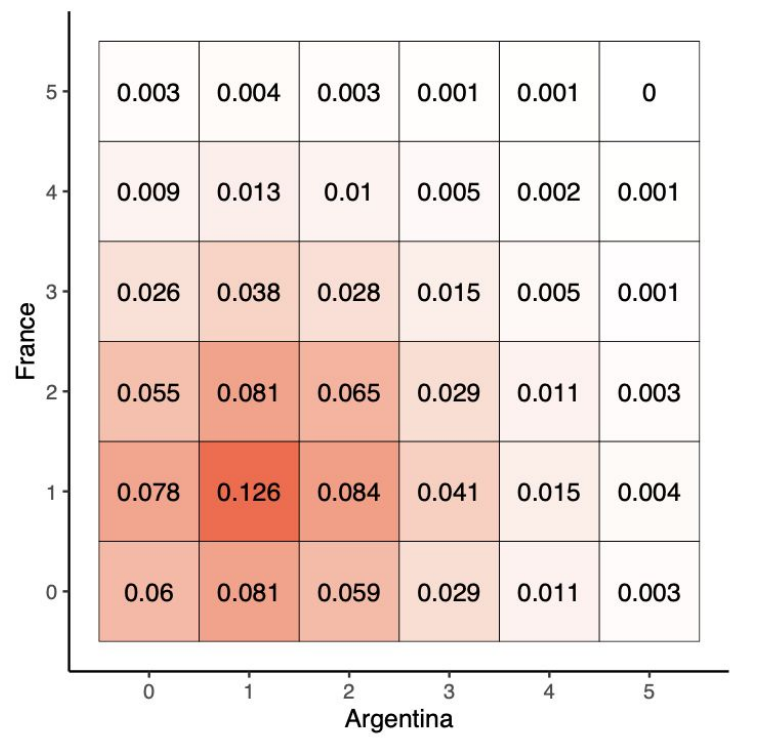
But how do we convert those numbers into match probabilities?

Creating the Match Result Matrix



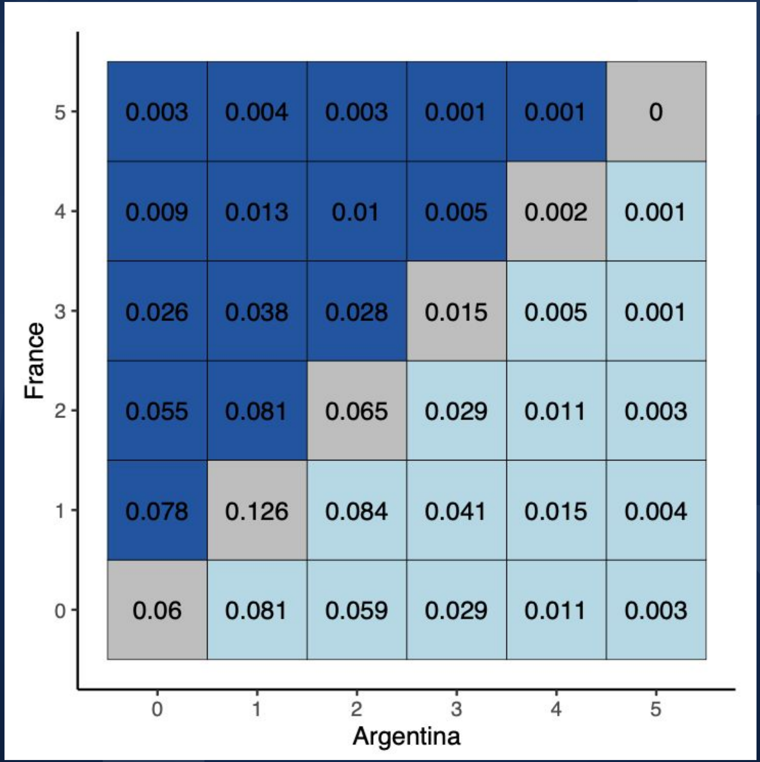
5	0.003	0.004	0.003	0.001	0.001	0
4	0.009	0.014	0.01	0.005	0.002	0.001
3	0.026	0.039	0.028	0.014	0.005	0.002
2	0.056	0.082	0.06	0.03	0.011	0.003
1	0.079	0.116	0.085	0.042	0.015	0.005
0	0.056	0.082	0.06	0.03	0.011	0.003
	0	1	2	3	4	5

Applying Draw Inflation Along the Diagonal



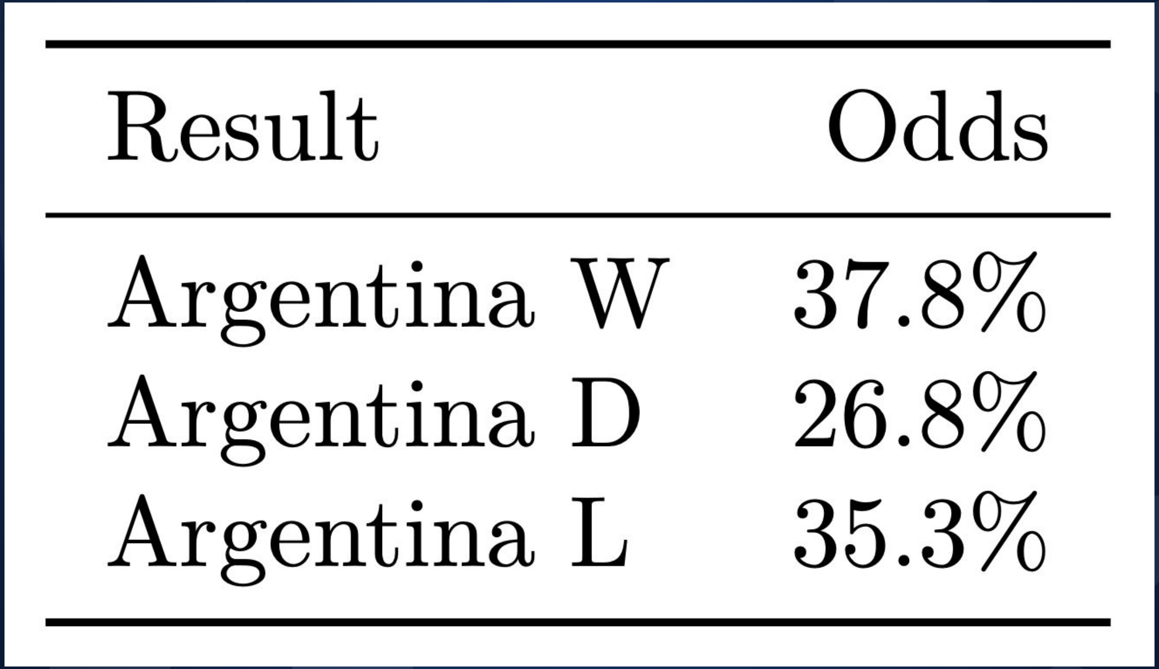
	0	1	2	3	4	5
5	0.003	0.004	0.003	0.001	0.001	0
4	0.009	0.013	0.01	0.005	0.002	0.001
3	0.026	0.038	0.028	0.015	0.005	0.001
2	0.055	0.081	0.065	0.029	0.011	0.003
1	0.078	0.126	0.084	0.041	0.015	0.004
0	0.06	0.081	0.059	0.029	0.011	0.003

Summing for Win/Loss/Draw Odds



France \ Argentina	0	1	2	3	4	5
5	0.003	0.004	0.003	0.001	0.001	0
4	0.009	0.013	0.01	0.005	0.002	0.001
3	0.026	0.038	0.028	0.015	0.005	0.001
2	0.055	0.081	0.065	0.029	0.011	0.003
1	0.078	0.126	0.084	0.041	0.015	0.004
0	0.06	0.081	0.059	0.029	0.011	0.003

Aggregating Final Predictions



Result	Odds
Argentina W	37.8%
Argentina D	26.8%
Argentina L	35.3%

4

Justification

Why Did We Pick The Data?

Team Data

- Necessary for how the team is doing and how they historically have done to predict match outcome
- Accounts for how a team compares to another team (who is better, worse, etc.) especially under certain significant conditions such as home advantage, win/loss rate of the season, etc.

Player Data

- Very indicative of how a team might do against another team based on player matchups which is more specific to the match than just a team's record in the season
- Using correlations and relationships between age, aggressiveness (yellow/red cards), player statistics, a team's potential against another team could be better calculated

Statistical Approach Flow

Goal Distribution

- Offensive/Defensive parameters based on team & player stats
- Initial parameters used for a team's goal distribution for the match

Joint Distribution

- Joint poisson distribution using the biased goal distributions to put both teams' distributions together

Home-Field Advantage

- Using goal distribution and then adding bias from home-field advantage (and perhaps, any other indicators

Draw inflation Distribution

- Used to account for the possibility of a draw – likely outcome of soccer matches based on historical records

Next Steps

- Proper Testing and Usage of Weights of the different predictors
  - How important is each variable in our dataset?
  - Do the player data or team data provide more significance for goal distribution?
- Building a Predictive Model
  - Expanding on the basic predictive model to account for intricate relationships in the data
  - Using the draw inflation in the model
  - Calculating how much the home-field indicator and draw inflation affects the match result matrix values