Results

Team Members:

Vivienne Hughes (ndg9tt)
Corinna Keum (pdx4tw)
Troy S. Meink (ked6na)
Brianna Seekford (bmx4af)

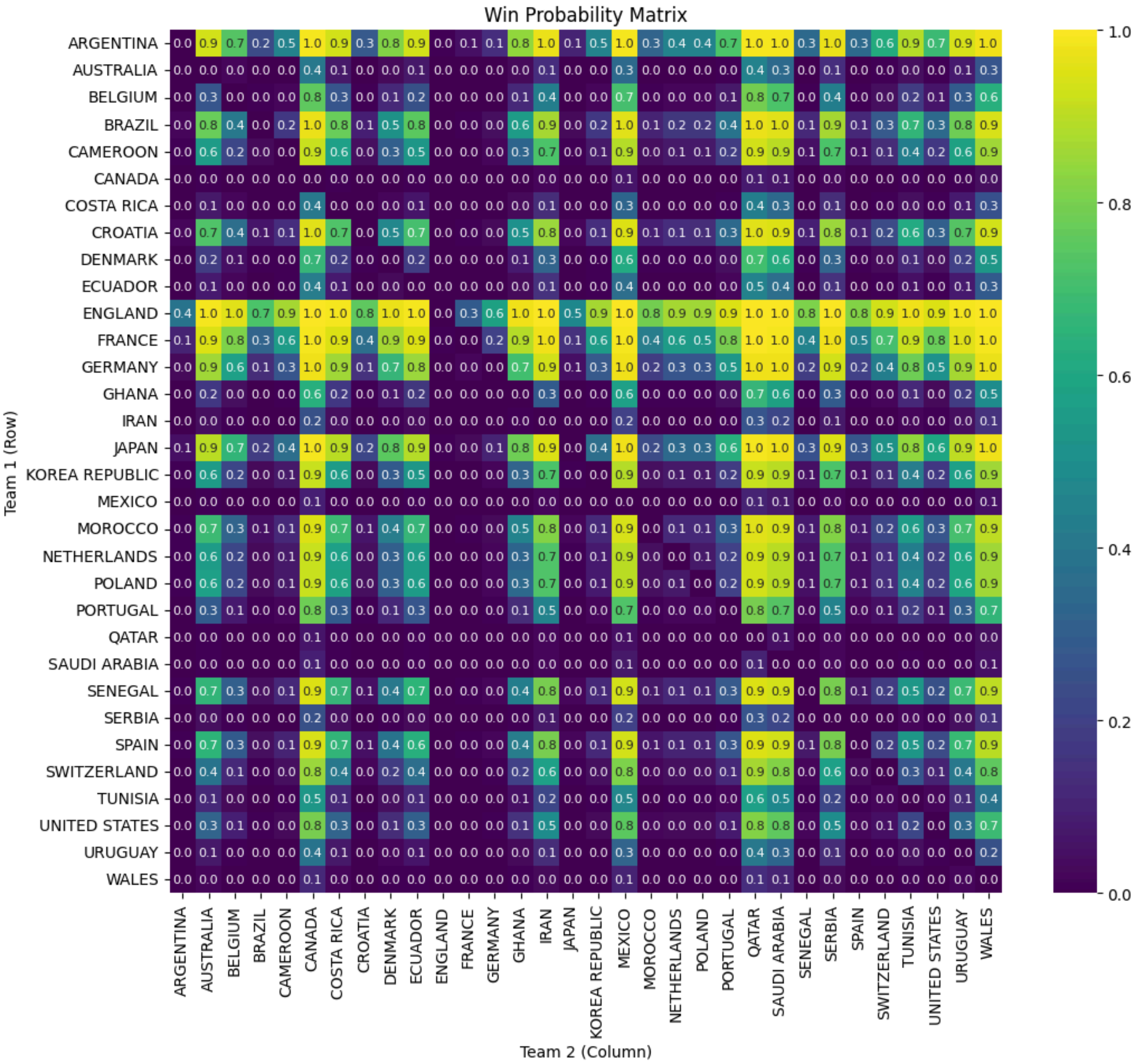
Based on a handful of unique datasets, our goal is to determine the final winner of the 2022 Men's World Cup in Qatar. A potential user base for our model, assuming it's effective, is sports bettors. 1.5 billion viewers tuned in to watch the world cup 2022 games [1]. Millions betting on online sites (like betonline.ag, mybookie, and bovada) and with family/friends, on who their favorite team is or home country. Creating a predictive model for the World Cup games can give sports bettors insight to potential betting strategies.

Our analysis focused primarily on two datasets, which were used to generate complementary models to predict the world cup outcome. First, the group stage results going into the 2022 Finals tournament was used in a logit-style regression to predict the match outcomes of individual pairs of teams. This was then used to run multiple simulated elimination brackets to determine the teams favored to win the overall tournament. Secondly, we used historical match data from 1993 to 2022. This dataset includes the home and away team, FIFA ranking, FIFA points, score of the match. For this dataset's model approach, two decision trees were used. One model was used to train on the likelihood of the home team or away team winning and then using metrics to rank the strongest teams. The other model generated all possible matches between the qualified teams for the 2022 games, to which teams are favorable to win. Our different models allowed us to contrast the influences of long-term team performance over many years to the extremely up-to-date data of a given World Cup's group stage results.

The primary analysis done with the 2022 group stage data involved filtering a large dataset down to the important statistics relevant to the outcome of any given match, including team possession, scoring attempts, goal preventions, etc. The differences in performance between winning and losing teams across all of these statistics was then used as the training data for a logit-style regression which used these statistics to predict individual match outcomes of individual pairs of teams.

From here, the average performance of each team through their group stage matches was used in conjunction with the regression to simulate various match-ups and tournament outcomes.

The matrix shown below represents the predicted outcome of every represented team against every other team. The number in each cell represents the likelihood the team in the corresponding row (Team 1) is to beat the one in the corresponding column (Team 2).



	ARGENTINA	AUSTRALIA	BELGIUM	BRAZIL	CAMEROON	CANADA	COSTA RICA	CROATIA	DENMARK	ECUADOR	ENGLAND	FRANCE	GERMANY	GHANA	IRAN	JAPAN	KOREA REPUBLIC	MEXICO	MOROCCO	NETHERLANDS	POLAND	PORTUGAL	QATAR	SAUDI ARABIA	SENEGAL	SERBIA	SPAIN	SWITZERLAND	TUNISIA	UNITED STATES	URUGUAY	WALES
ARGENTINA	0.0	0.9	0.7	0.2	0.5	1.0	0.9	0.3	0.8	0.9	0.0	0.1	0.1	0.8	1.0	0.1	0.5	1.0	0.3	0.4	0.4	0.7	1.0	1.0	0.3	1.0	0.3	0.6	0.9	0.7	0.9	1.0
AUSTRALIA	0.0	0.0	0.0	0.0	0.0	0.4	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.4	0.3	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.3
BELGIUM	0.0	0.3	0.0	0.0	0.0	0.8	0.3	0.0	0.1	0.2	0.0	0.0	0.0	0.1	0.4	0.0	0.0	0.7	0.0	0.0	0.0	0.1	0.8	0.7	0.0	0.4	0.0	0.0	0.2	0.1	0.3	0.6
BRAZIL	0.0	0.8	0.4	0.0	0.2	1.0	0.8	0.1	0.5	0.8	0.0	0.0	0.0	0.6	0.9	0.0	0.2	1.0	0.1	0.2	0.2	0.4	1.0	1.0	0.1	0.9	0.1	0.3	0.7	0.3	0.8	0.9
CAMEROON	0.0	0.6	0.2	0.0	0.0	0.9	0.6	0.0	0.3	0.5	0.0	0.0	0.0	0.3	0.7	0.0	0.1	0.9	0.0	0.1	0.1	0.2	0.9	0.9	0.1	0.7	0.1	0.1	0.4	0.2	0.6	0.9
CANADA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
COSTA RICA	0.0	0.1	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.4	0.3	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.3
CROATIA	0.0	0.7	0.4	0.1	0.1	1.0	0.7	0.0	0.5	0.7	0.0	0.0	0.0	0.5	0.8	0.0	0.1	0.9	0.1	0.1	0.1	0.3	1.0	0.9	0.1	0.8	0.1	0.2	0.6	0.3	0.7	0.9
DENMARK	0.0	0.2	0.1	0.0	0.0	0.7	0.2	0.0	0.0	0.2	0.0	0.0	0.0	0.1	0.3	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.7	0.6	0.0	0.3	0.0	0.0	0.1	0.0	0.2	0.5
ECUADOR	0.0	0.1	0.0	0.0	0.0	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.5	0.4	0.0	0.1	0.0	0.0	0.1	0.0	0.1	0.3
ENGLAND	0.4	1.0	1.0	0.7	0.9	1.0	1.0	0.8	1.0	1.0	0.0	0.3	0.6	1.0	1.0	0.5	0.9	1.0	0.8	0.9	0.9	0.9	1.0	1.0	0.8	1.0	0.8	0.9	1.0	0.9	1.0	1.0
FRANCE	0.1	0.9	0.8	0.3	0.6	1.0	0.9	0.4	0.9	0.9	0.0	0.0	0.2	0.9	1.0	0.1	0.6	1.0	0.4	0.6	0.5	0.8	1.0	1.0	0.4	1.0	0.5	0.7	0.9	0.8	1.0	1.0
GERMANY	0.0	0.9	0.6	0.1	0.3	1.0	0.9	0.1	0.7	0.8	0.0	0.0	0.0	0.7	0.9	0.1	0.3	1.0	0.2	0.3	0.3	0.5	1.0	1.0	0.2	0.9	0.2	0.4	0.8	0.5	0.9	1.0
GHANA	0.0	0.2	0.0	0.0	0.0	0.6	0.2	0.0	0.1	0.2	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.7	0.6	0.0	0.3	0.0	0.0	0.1	0.0	0.2	0.5
IRAN	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.3	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1
JAPAN	0.1	0.9	0.7	0.2	0.4	1.0	0.9	0.2	0.8	0.9	0.0	0.0	0.1	0.8	0.9	0.0	0.4	1.0	0.2	0.3	0.3	0.6	1.0	1.0	0.3	0.9	0.3	0.5	0.8	0.6	0.9	1.0
KOREA REPUBLIC	0.0	0.6	0.2	0.0	0.1	0.9	0.6	0.0	0.3	0.5	0.0	0.0	0.0	0.3	0.7	0.0	0.0	0.9	0.0	0.1	0.1	0.2	0.9	0.9	0.1	0.7	0.1	0.1	0.4	0.2	0.6	0.9
MEXICO	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
MOROCCO	0.0	0.7	0.3	0.1	0.1	0.9	0.7	0.1	0.4	0.7	0.0	0.0	0.0	0.5	0.8	0.0	0.1	0.9	0.0	0.1	0.1	0.3	1.0	0.9	0.1	0.8	0.1	0.2	0.6	0.3	0.7	0.9
NETHERLANDS	0.0	0.6	0.2	0.0	0.1	0.9	0.6	0.0	0.3	0.6	0.0	0.0	0.0	0.3	0.7	0.0	0.1	0.9	0.0	0.0	0.1	0.2	0.9	0.9	0.1	0.7	0.1	0.1	0.4	0.2	0.6	0.9
POLAND	0.0	0.6	0.2	0.0	0.1	0.9	0.6	0.0	0.3	0.6	0.0	0.0	0.0	0.3	0.7	0.0	0.1	0.9	0.0	0.1	0.0	0.2	0.9	0.9	0.1	0.7	0.1	0.1	0.4	0.2	0.6	0.9
PORTUGAL	0.0	0.3	0.1	0.0	0.0	0.8	0.3	0.0	0.1	0.3	0.0	0.0	0.0	0.1	0.5	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.8	0.7	0.0	0.5	0.0	0.1	0.2	0.1	0.3	0.7
QATAR	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SAUDI ARABIA	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
SENEGAL	0.0	0.7	0.3	0.0	0.1	0.9	0.7	0.1	0.4	0.7	0.0	0.0	0.0	0.4	0.8	0.0	0.1	0.9	0.1	0.1	0.1	0.3	0.9	0.9	0.0	0.8	0.1	0.2	0.5	0.2	0.7	0.9
SERBIA	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
SPAIN	0.0	0.7	0.3	0.0	0.1	0.9	0.7	0.1	0.4	0.6	0.0	0.0	0.0	0.4	0.8	0.0	0.1	0.9	0.1	0.1	0.1	0.3	0.9	0.9	0.1	0.8	0.0	0.2	0.5	0.2	0.7	0.9
SWITZERLAND	0.0	0.4	0.1	0.0	0.0	0.8	0.4	0.0	0.2	0.4	0.0	0.0	0.0	0.2	0.6	0.0	0.0	0.8	0.0	0.0	0.0	0.1	0.9	0.8	0.0	0.6	0.0	0.0	0.3	0.1	0.4	0.8
TUNISIA	0.0	0.1	0.0	0.0	0.0	0.5	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.2	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.6	0.5	0.0	0.2	0.0	0.0	0.0	0.0	0.1	0.4
UNITED STATES	0.0	0.3	0.1	0.0	0.0	0.8	0.3	0.0	0.1	0.3	0.0	0.0	0.0	0.1	0.5	0.0	0.0	0.8	0.0	0.0	0.0	0.1	0.8	0.8	0.0	0.5	0.0	0.1	0.2	0.0	0.3	0.7
URUGUAY	0.0	0.1	0.0	0.0	0.0	0.4	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.4	0.3	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.2
WALES	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 1. Win probability confusion matrix

This graphic was essentially used as a sanity check to ensure the regression was making reasonable sense. Many teams which were expected to perform well by the model, such as England, France, or Germany, were also favored by analysts before the world cup even began.

Next, randomly seeded tournaments were run with the model determining the outcome of each individual match. These were used in aggregate to determine the teams most favored by the model to win the entire tournament.

Figure 2 shows the outcome of one of these simulations with 100 randomly seeded tournaments.

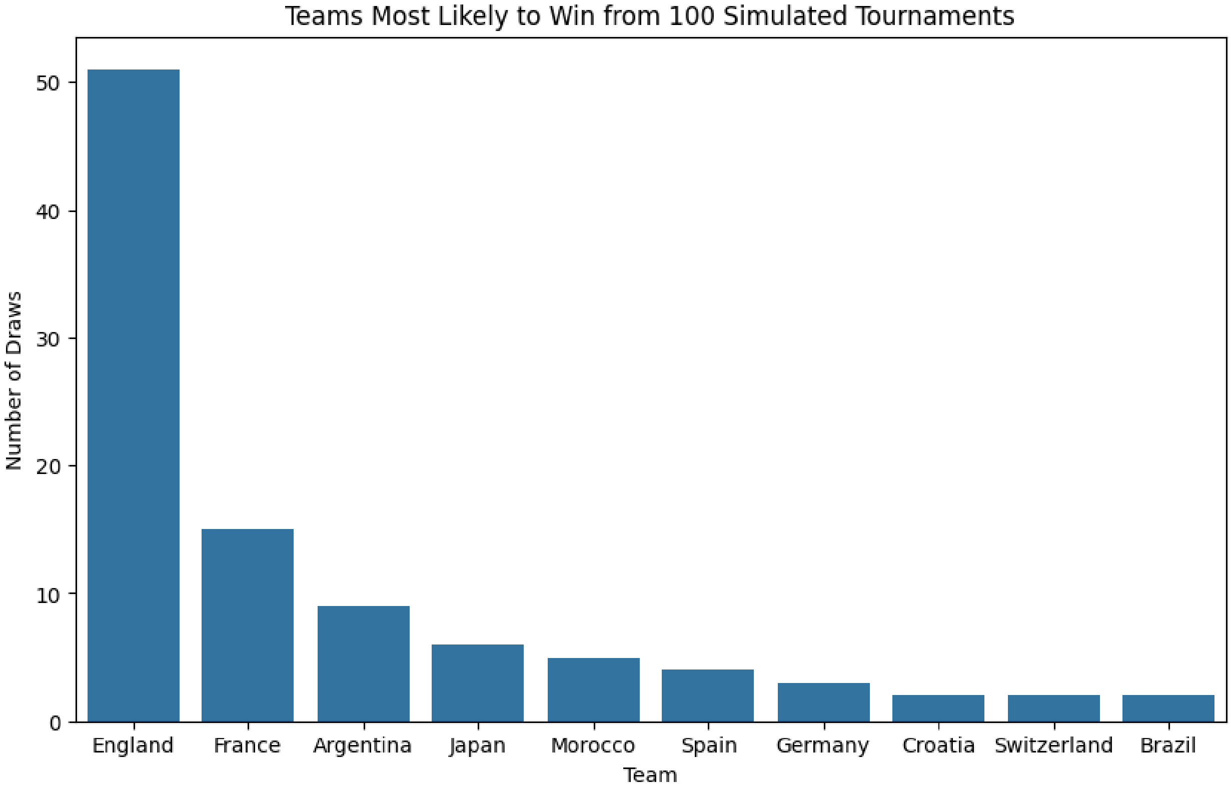


Figure 2. Winners from 100 simulated tournaments.

As is clearly shown England is heavily favored to win, by a factor of almost 2-to-1 when compared to the rest of the field. The skewed outcome of many of these simulations is likely due to the small number of matches used in training the model. This is why we ran similar regressions on historical data over a number of years.

Some attempts were made to lower the confidence the model had that England would win. This included scaling the model coefficients and intercept by a number less than 1 to reduce how much it favored high performing teams. However, these had little effect on the outcome of the simulations, indicating that this behavior was a core aspect of how the model fitted the data.

Next we will compare the results to our second dataset analysis. The first model used to predict the top strongest teams calculated historical statistics of each team's FIFA points, ranks, and win and then ranked the teams based on the aggregated performance metrics. The top 10 predicted teams of Total FIFA points, FIFA rank, and Total Wins are shown in Figure 3.

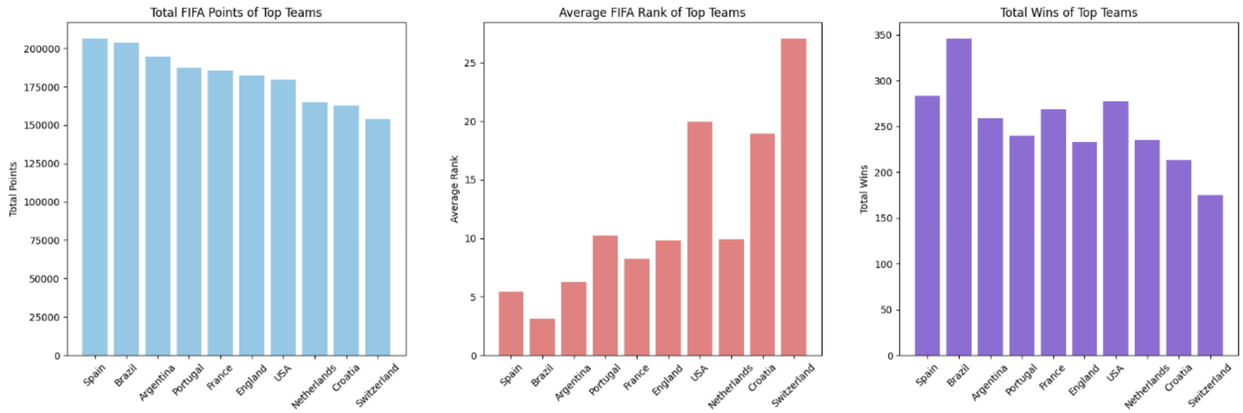


Figure 3. Three Bar Graphs of Top 10 Strongest Teams based on Total FIFA Points, FIFA Rank, and Total Wins

Our model showed that the top 3 predicted teams are Spain (Total FIFA Points: 206508, Average FIFA Rank: 5.42, Wins: 283), Brazil (Total FIFA Points: 203947, Average FIFA Rank: 3.12, Wins: 346), and Argentina (Total FIFA Points: 194845, Average FIFA Rank: 6.30, Wins: 259). This model has an accuracy of 99.12% in predicting the outcome of Home team and Away team wins. This model however is not does not simulate matches between all of the qualified teams, just for the teams that have already taken place. Thus, we created a second Decision Tree model that would generate matches between each qualified team to predict the winner, by using itertools.permutation and matching and scaling the rank and points difference of the teams to determine the outcome of the match. This model focuses on specific matchups making it better for specific comparisons and a better simulation of the tournament.

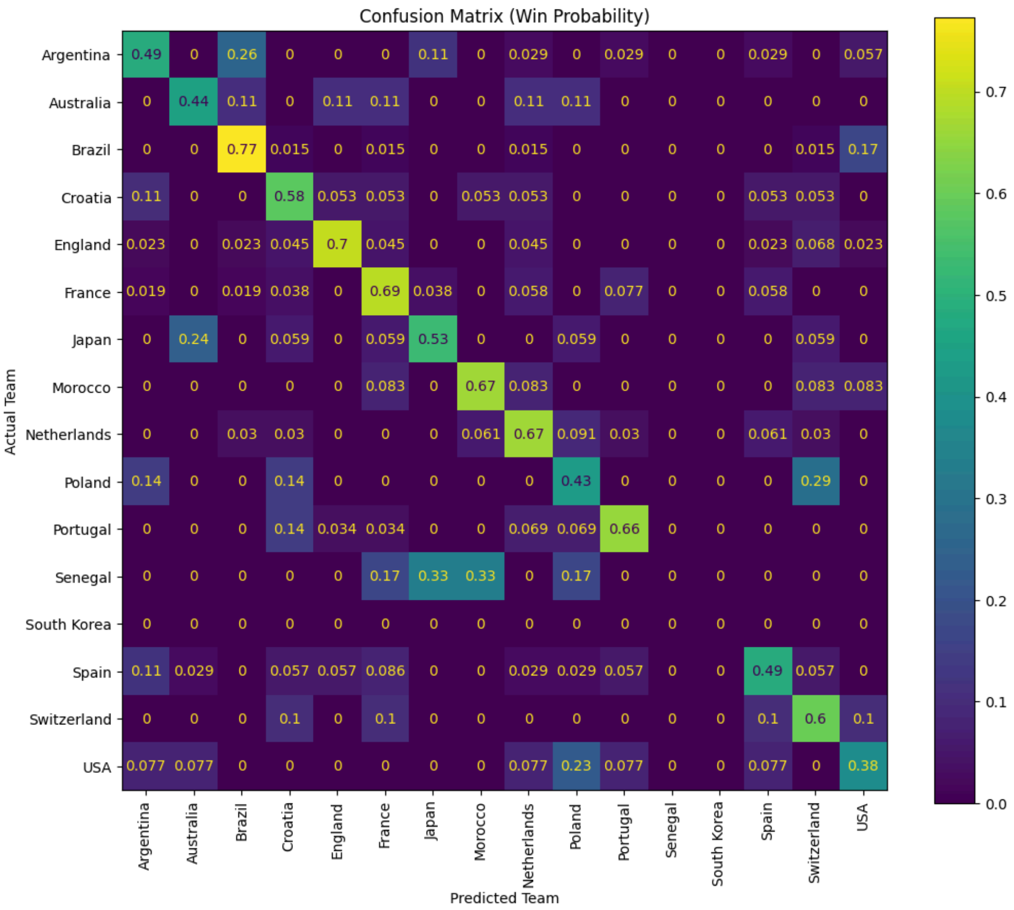


Figure 4. Win probability confusion matrix (Model 2)

The model's confusion matrix is shown in Figure 4. The diagonal values represent the correct predictions of each team. Brazil has the highest diagonal value of 0.77, which suggests that Brazil is often predicted correctly as a winner. England has the second strongest diagonal value of 0.70, followed by Morocco's accuracy of 0.67. Countries that have a lower accuracy because the model shows skepticism in some of the predicted results when Morocco had won. This is shown in the off-diagonal cells of the confusion matrix in Figure 4 because cells outside of the diagonal represent when there is a misclassification. The model's overall accuracy is 73.19%. This lower accuracy is because of having a simpler feature set (that only compares the features: rank and point difference) which limits the amount of information the model has to make the predictions.

The results of this model is shown in Figure 5, a bar graph of the predicted wins for each team for this model. The biggest comparison between this model and the first model discussed is the lack of dominance of England in this simulation. This difference might be from the model is the regression, which takes sequences into account. Both Figure 2, Figure 3, and Figure 5 show that France and Argentina are very strong teams in both predictions.

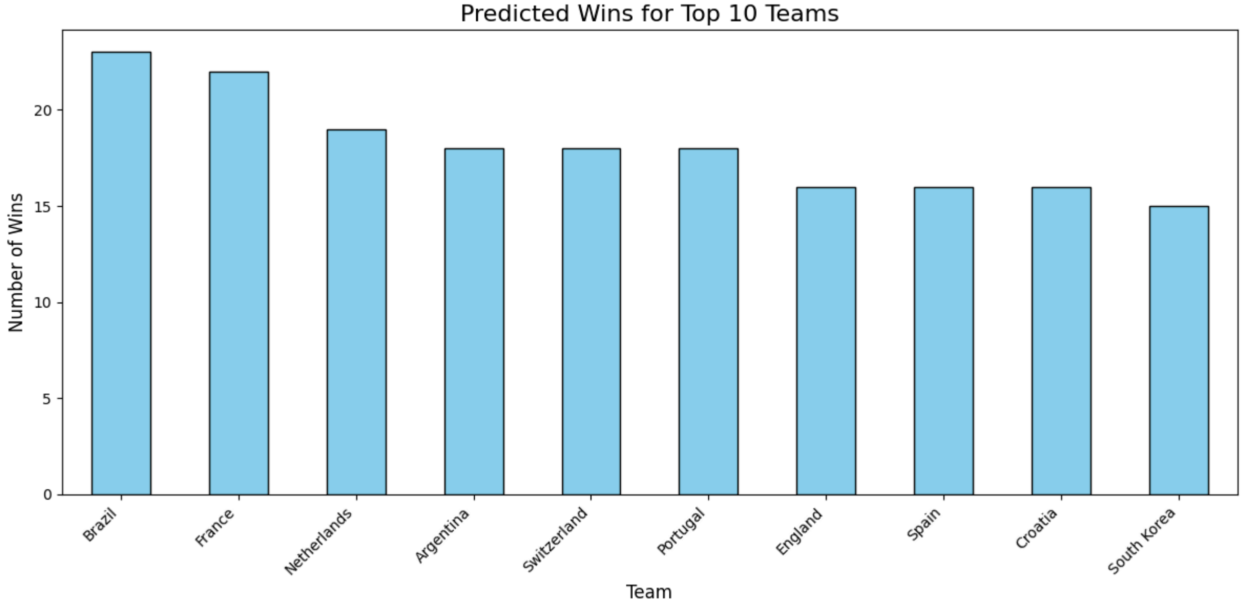


Figure 5. Predicted Wins for the Top 10 Teams

References

[1] “Inside FIFA,” in numbers, https://inside.fifa.com/fifa-world-cup-qatar-2022-in-numbers (accessed Dec. 8, 2024).