DS 3001 Final Paper

December 14, 2024

Advisor:

Terence Johnson, School of Data Science

Team Members:

Vivienne Hughes, Corinna Keum, Troy S. Meink, Brianna Seekford

Abstract

Based on a handful of unique datasets, our goal was 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. Given a model trained from the 2022 World Cup, sports bettors could use it to bet on upcoming tournaments, as well as individual match outcomes.

Our models' statistics can benefit more than the average World Cup fan, but also sport analysts (the model may provide data insights to improve commentary), coaches of the teams (refine strategies based on statistics), and anyone interested in data science.

Because we drew from a number of different sources, users of our models would have flexibility in the data they use to make predictions as well as the way in which those predictions are made. Additionally, the models were developed simply, both in terms of the process to input data to make predictions, as well as how results are presented to the user. This would allow them to be useful for both data scientists and those less experienced with programming.

Introduction

The FIFA Men's World Cup is one of the most anticipated global sporting events in the world, captivating billions of viewers and generating widespread enthusiasm. The 2022 tournament held in Qatar drew an audience of over 1.5 billion with millions engaging not just as spectators but also as active participants in sports betting, on platforms like betonline.ag, my bookie, and bovada, or in informal pools with friends and family [1]. Predicting match outcomes for games like this can be just for fun, but it creates an opportunity for data scientists looking for a deeper understanding of team performance. Data scientists can apply machine learning and data-driven methods to enhance insights. The primary goal of this study was to develop predictive models capable of determining the winner of the 2022 Men's World Cup in Qatar. This project's analysis seeks to provide a foundation for knowledge of the outcome of the tournament and individual matches.

Our project began with initial data collection for historical data on the teams playing in the 2022 games. We obtained our data from the Kaggle website. We went through the steps of wrangling our two datasets we are using for this project. This step was important to eliminate unnecessary data and to deal with null or missing values. We decided to split the datasets amongst the group to create different models to compare from. The first model's modeling approach was using logistic regression to simulate elimination brackets and predict the likelihood of individual match outcomes. This model focuses on the specific matchups which is significant in determining the outcome of the games because it makes it better for specific comparisons and a better simulation of the tournament. The second dataset's modeling approach was to employ a decision tree to generate predictions. The second dataset had two directions for the predictions, one was to calculate the historical statistics and rank teams on aggregated performance metrics.

This model emphasizes the historical wins and loses of a specific team, which the prediction is a clear indicator of the calculated historical statistics of each team. And the other decision tree model generated matches between each qualified team to predict the winner and scaling the rank. This decision tree model was significant in simulating a tournament because of its random generating matches. The dual dataset approach offered a comprehensive perspective of how use of different datasets and modeling approaches can cause different predictions.

Our model methodology was designed to process key statistics, like the possession rates, scoring attempts, and goal preventions, to identify performance differences between teams. This information was used to simulate tournament outcomes, validate predictions highlighting potential champions, and emphasized key contenders to be the winner of the World Cup (like England, Brazil, and Argentina). Additionally, we will identify the strengths, limitations, and accuracy of different modeling approaches, logistic regression and decision tree.

This paper provides our steps and reasoning for the development of our project from the anticipated failures of our model and data to the accuracy and uncertainties in our predictions evaluation shown through the confusion matrix. Our project provides a comprehensive comparison of the output and predictions of our logarithmic regression and decision tree models through figures showing the top predicting figures of the different models. This paper includes our trials and errors from deciding which datasets to use to the results of the predictions. Lastly, we will end on a conclusion that compiles our findings with how we could use our findings for the future predictions and projects. Our predictive models can offer valuable tools for enhancing the World Cup experience, for both casual fans and advanced analysts.

Data

We started with four datasets which covered different aspects of the 2022 Men's World Cup. After conducting some data wrangling and EDA, it was realized that two datasets would not be viable to move forward with so we only did our primary analyses on the remaining two.

Corinna's data was historical games from 1993 to 2022. This dataset's observations were from individual matches, with a home team and away team from friendly tournaments to World Cup qualification matches. The two teams were combined in one variable for the specific match data, so to specify the specific team statistics, there is an indication, like "home_team_rank" and "away_team_rank". This CVS had fields for both team's goalkeeper score, midfield_offence score,_midfield_score, mean_defense_score, mean_offense_score, and mean_midfield_score. However these columns had too many null and missing values that the data set for analyzing the specific scores of the positions were not sufficient enough. We deleted all the matches before 1993 because we assumed that only the last few World Cups as players on the team will likely be the same, so this will show who performs better and usually makes it farther. Even though there were many fields that needed to be deleted because of the inadequacy of the missing values, we were still able to keep the FIFA ranking, FIFA points, score of the match (for each of the home and away teams). 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 from the FIFA ranking, FIFA points, and score of the match.

Vivienne's data contained group stage data for the 2022 World Cup. There were tens of categorical and numerical variables for each team and match. First, many of these variables which were not match-related were removed. These included the city each team was from and similar metrics.

From there, some preliminary analyses were done to get an understanding of how some of the different variables could be used to gauge the performance of each team. This included comparing how well certain variables corresponded with win or loss match outcomes for each team. An interesting result from this was that there wasn't significant correlation between ball possession and winning match outcomes. This indicates that whether or not a team maintained possession of the ball for a large percentage of the match, didn't mean they performed better than teams that didn't. Conversely, high scoring attempts, even if they're missed, have a high correlation with winning. This shows that more proactive teams that take more shots on the goal tend to win more frequently than those making less attempts.

Because Vivienne's data was extremely clean as well as relevant to the 2022 finals matches, we decided to move forward with it. This group stage data would provide a useful comparison to Corinna's historical data.

Brianna's data consisted of statistics from individual matches during the 2022 season. It included details on predicted scores, the Soccer Power Index, the actual outcomes of the matches, and other information surrounding the matches. The variables were separated into team1 and team2 to allow for quick comparison. There were many variables however that complicated the set such as what types of final game goals were scored and where exactly on the field they were scored from. It was a bit hard to understand what some of the variables were symbolizing and there was no file to explain on the site where the data was collected.

Troy's data had analyst predictions for the outcome of the Men's 2022 World Cup. This included predictions for the group stage performance of every time, as well as the likelihood they would reach various points in the finals bracket.

The data were extremely clean with no null or missing values and clear descriptions. Some important variables Troy looked into were the simulated number of wins/losses, goals scored for/against and a team's projected final position.

Because the data was so well-formatted, it would have been easy to work with. However, it was decided this data would not be as good a fit for our project as some of the other historical datasets. This is because we would be using predictions of the World Cup outcome to make our own predictions as to that outcome.

Methods

As discussed in the previous section, we narrowed down the scope of our project from four to two datasets. This was done based on the comparatively low quality of the two predication-based datasets compared to the historical and group stage data. These would provide an interesting comparison between the long term performance of teams and their most recent performance in the 2022 group stage. From here, work was done to form the basis of the rest of our analysis.

The historical data from individual World Cup matches would be the unit of observation for our study. Based on match results from the 2022 group stage and the historical data from previous World Cup years, two models could be used to make individual match outcome predictions.

Supervised learning models were used for both datasets. A Logit-style regression was done with the group stage data and two decision trees were used with the historical data. In both cases there were match results (y-values) that the models were fitted to. Because the predictions of both models were essentially binary (will a team win or lose?), they would use classification, not regression-based analyses.

As discussed previously, a logistic regression was done on the group stage data and decision trees were used on the historical data. These were chosen because they were most conducive to the variables and data types in each dataset. These would allow us to fit the models based on previous team performance and match results. From here, simulated tournaments could be run to predict the outcome of finals matches.

The predictions of our models could then be compared to both team rankings going into the 2022 season, as well as the actual results of the 2022 World Cup. This benefit of being able to

compare to actual season results was why we chose to work with data from a previous World Cup rather than a future one. Success for our models meant that they accurately captured the overall performance of the World cup teams. Given the limited scope of our analyses, we decided it was too ambitious to attempt predictions that closely matched the actual final results. However, we wanted our models to pick up on which teams were high performers going into the World Cup and which ones would do well in the finals bracket.

Given the nature of elimination brackets, we anticipated that it would be increasingly difficult to predict winners the farther we got into the finals matches because if a model made an incorrect prediction in one match it would propagate through the entire bracket. This was the primary weakness we were concerned about going into the analysis. However, running multiple simulated tournaments would allow us to see the average performance of teams, rather than outliers that overperformed in individual brackets.

Both the historical and group stage date came from Kaggle so it was extremely clean and didn't require much wrangling. Much of the data was numerical, the handful of categorical variables we had were one-hot encoded on a case-by-case basis. Some preliminary analyses indicated that there wasn't much correlation between any of the variables.

We were able to use a number of different methods to present the results of our analyses. Win probability matrices were an excellent way to represent the predicted strength of *every* team against each other. These allowed for quick assessments to be made on the likely performance of each team based on how likely they were to win against the others. Histograms were also a great way to display the top ranked teams by the models. This included those most likely to win the tournament based on simulated finals matches and points rankings. The specifics of how these results are determined and presented are discussed in the next section.

Results

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).

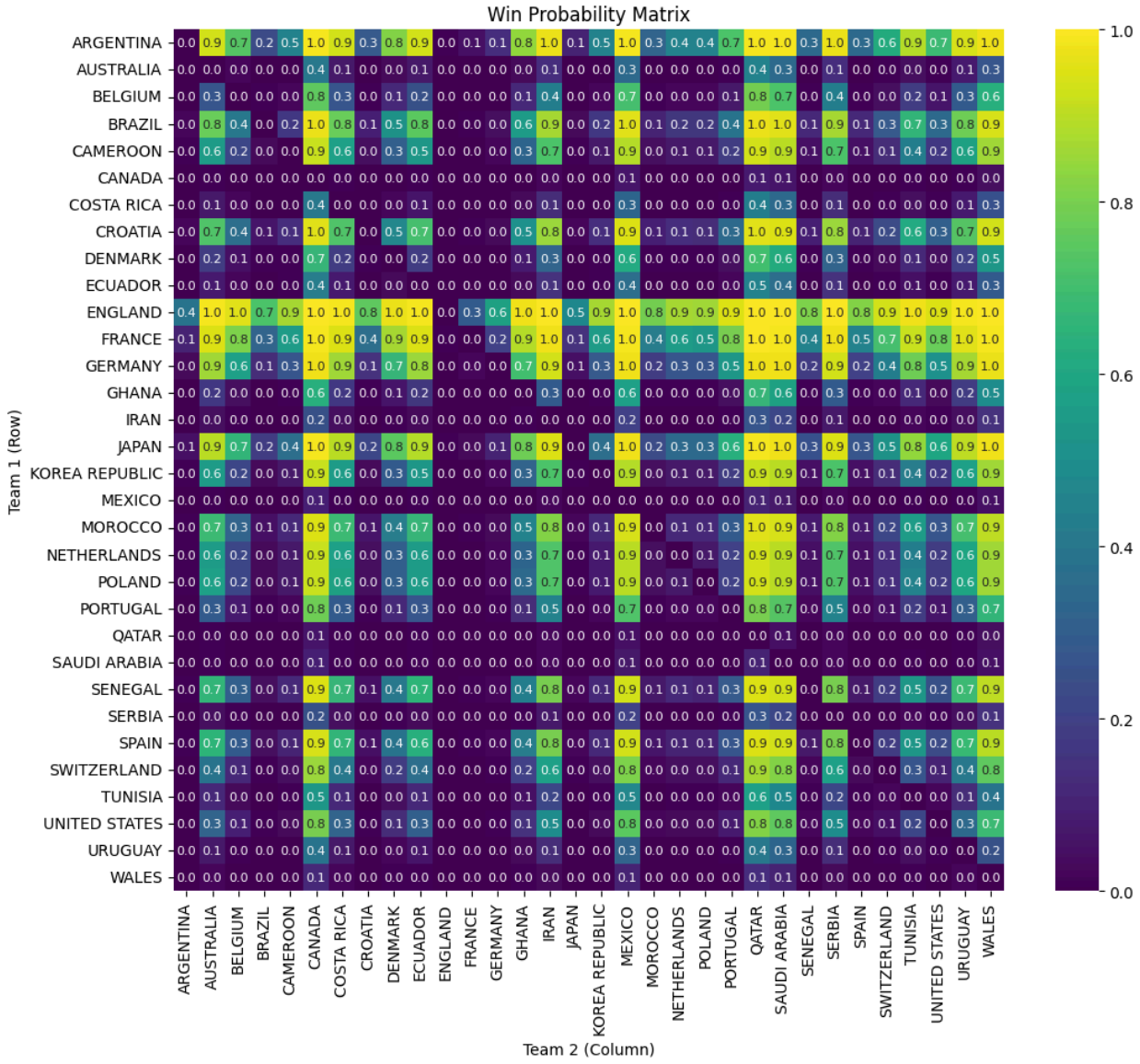

Figure 1. Win probability confusion matrix

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.

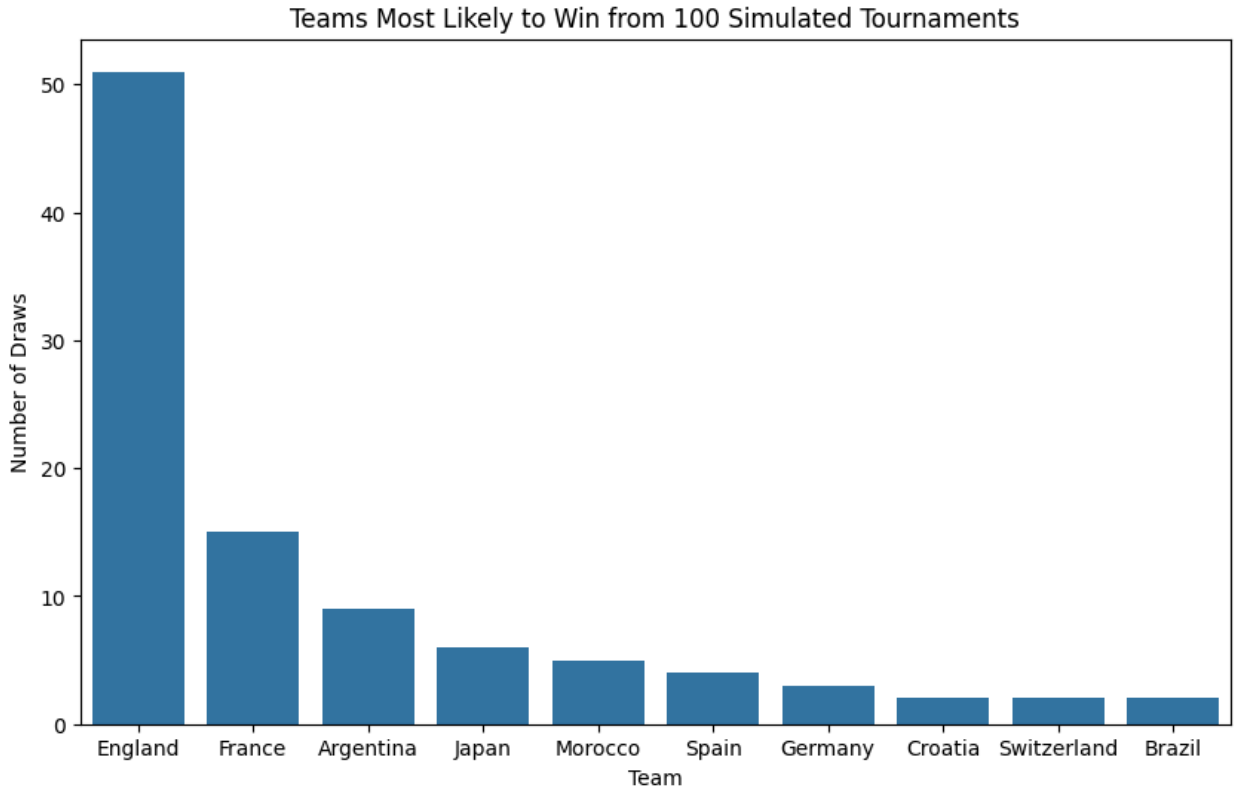

Figure 2. Winners from 100 simulated tournaments.

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.

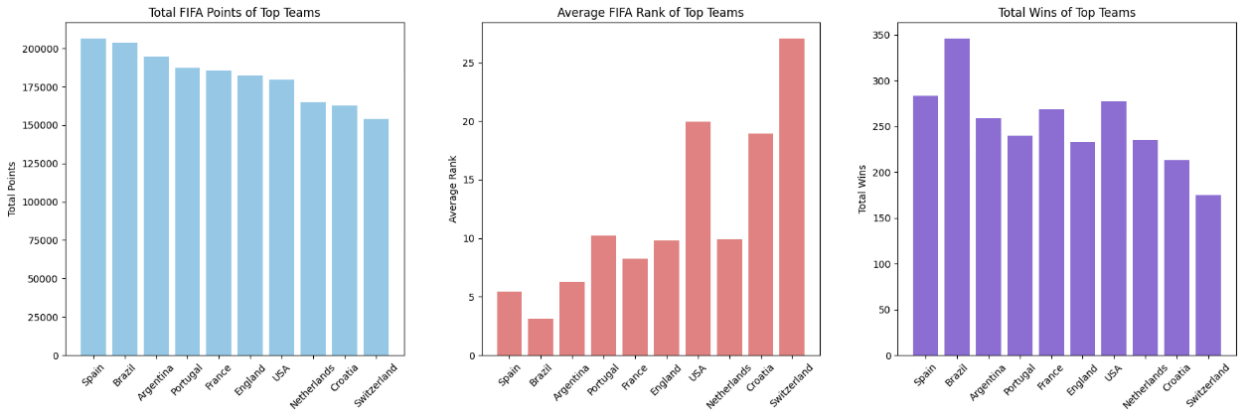

Figure 3. Three Bar Graphs of Top 10 Strongest Teams based on Total FIFA Points, FIFA Rank, and Total Wins

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.

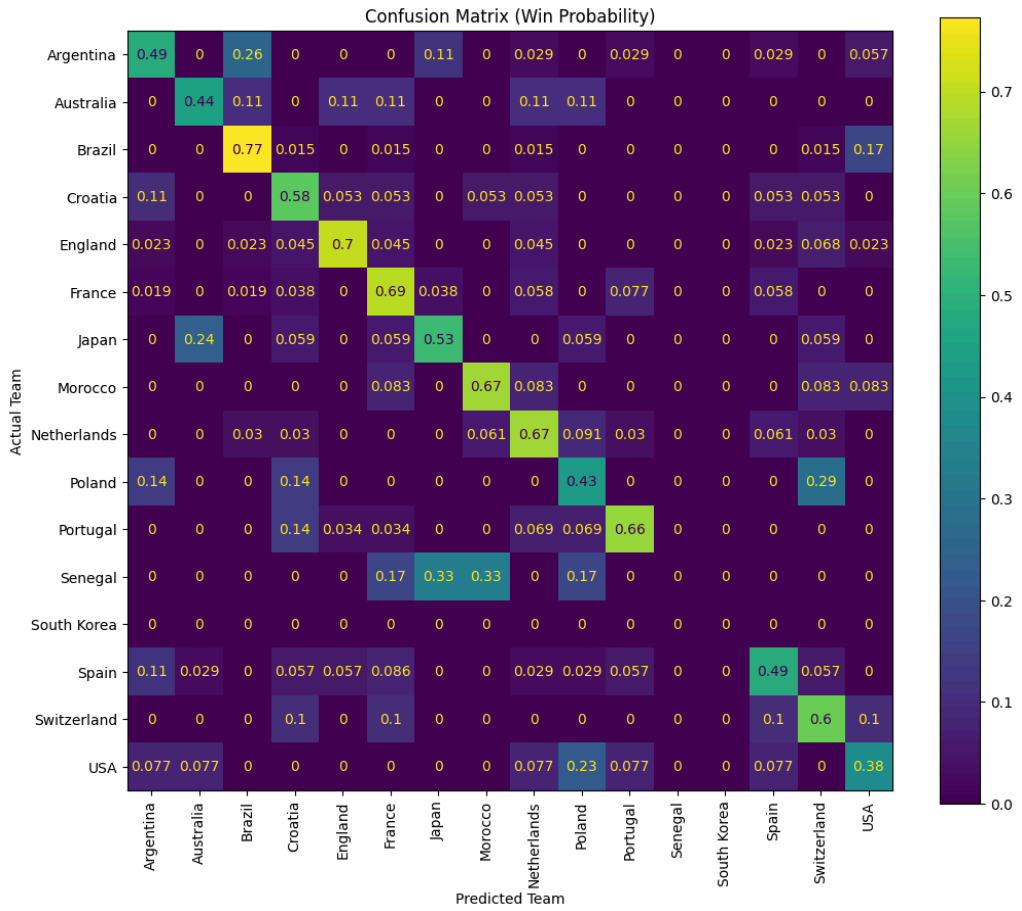

Figure 4. Win probability confusion matrix (Model 2)

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.

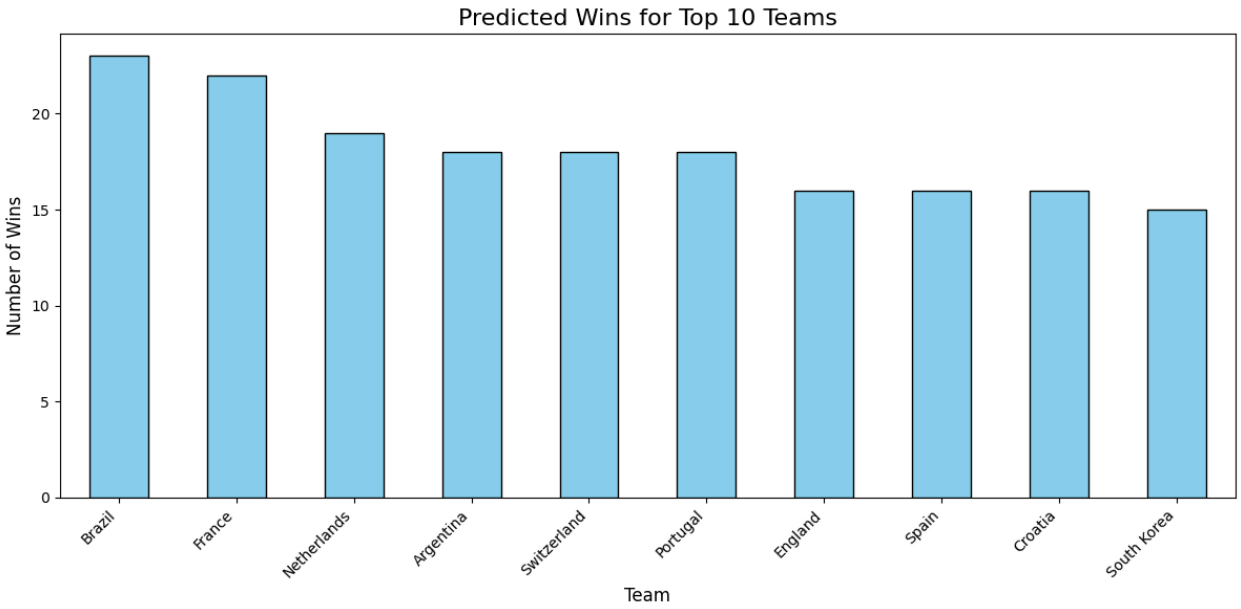

Figure 5. Predicted Wins for the Top 10 Teams

Conclusion

Introduction

This project aimed to find the best program to be able to predict the winner of the World Cup. Through a combination of logistic regression and decision tree models, we were able to help analyze data to lead us to a predicted winner.

The first model that we used for our analysis used 2022 group stage data that allowed us to look at individual matches and their results. Our second focused more on historical data from previous years which helped us to calculate the team strengths with focuses on variables such as FIFA points, ranking, and overall wins. By using a regression model, we were able to identify England as a strong team as a predicted winner. With our decision tree model, however, we were able to analyze more variables and find more common trends. This led to Brazil and Argentina as two of the most favored contenders.

Model Insights and Performance

With England, Brazil, and Argentina being our top contenders, we are able to compare this to real world findings. These teams consistently do well and are top in the league which legitimizes our findings. However, when using limited data, there is bias than can occur. With our logit-style regression model, the variable ranking was taken into consideration a great deal. When upsets are very common in the final rounds of pro sports championships, it is important to look at other factors as matches occur. Our decision tree model helped bring in more variables for analysis. This led to more diverse predictions and had a higher accuracy for predicted outcomes.

Challenges and Limitations

There were several challenges that we faced throughout the project. When looking at a sport like soccer, where there are so many variables that can go into what makes a winning team, it is hard to find data that can help predict a World Cup winner. There are also many factors that are unpredictable that can cause upsets. Factors such as injured players or weather are examples of variables we were not able to include in our model, but could have been used to improve the accuracy of the predictions our models had.

References

[1] “Inside FIFA,” in numbers,

https://inside.fifa.com/fifa-world-cup-qatar-2022-in-numbers (accessed Dec. 8, 2024).