2022 FIFA World Cup Predictor

Abstract:

Being a huge football fan, I wanted to make a model that would predict the winners of the next edition of the FIFA world cup which is being played in 2022 in Qatar. I modeled which team would win against another by having two classes that each team would take: home team and away team. I divided the teams playing the world cup into groups of 4 teams as drawn by FIFA, and within each group every team plays every other team once. In the group stages, I assigned 3 points to each team for a win, 1 point for a draw, and 0 points for a loss. The top two teams from each group with the highest points progress to the knockout stage. In the knockout stage whichever team wins their game progresses to the next round, while the losing team is eliminated. I simulated the results of each game by picking the winner as the team with the highest probability of winning according to the data which was the performance of each team playing the 2022 world cup in international matches from the year 2000 onwards as well as the latest FIFA men's rankings. The model predicted England and Brazil reaching the finals, with Brazil being crowned champions. These predictions are consistent with predictions generated by a lot of reputed betting and sport analytic companies which all see Brazil taking the cup home.

Introduction:

High profile sport tournaments capture the eyes of millions of people who eagerly await their team to win. Since football is nothing short of a religion in most parts of the world, the highest level of international competition, the FIFA world cup, has the power to make nations as whole wonder which team will take the world cup home this time. I am one such person, wondering which team will go all the way in the next edition of the world cup which will be played in 2022, hoping to leverage my machine learning skills to answer this billion dollar question. Correctly predicting the winners of the world cup has major implications in the worlds of betting and sports analytics. To put its magnitude into perspective, close to $155 billion was bet on different matches of the 2018 FIFA world cup and various players who turned out to be stars for their teams got contracts worth millions of dollars with different clubs.

The model had two inputs: the latest men's rankings published by FIFA and results of all international football matches that were played from the year 1872 to 2022 which I found on Kaggle. I edited the file containing the results of all international matches so that results prior to the year 2000 were filtered out and not used in the prediction.

I then used a SVM to train the model for the group stages and logistic regression to train the model for the knockout stages of the tournament since the testing data indicated that SVM and logistic regression had the highest accuracies out of all the models that were considered.

According to FIFA rules, a game can only end as a draw in the group stage but a game is not allowed to end in a draw in the knockout stage. The model outputs its prediction (win, loss, or draw) for each game that will be played in the 2022 edition of the FIFA world cup along with the probabilities of either team winning or the game ending in a draw (only in the group stages). The output of each match is then synthesized so that the results of the group stages are stored in tables such that there is one table representing the results of each group, and in a elimination style bracket for the knockout stages to visually display the results of the round of 16, quarterfinals, semifinals, and the final itself.

Dataset and Features:

I utilized multiple datasets–game results and fifa rankings–that were combined together to create a comprehensive data set for the assessment of the 2022 Fifa World Cup. Specifically I used 17,317 total international games dating back to 2000. The included information about the games was; the teams competing, who has home field advantage, the scores of the game, date played, and location of the game. From this dataset I created a dataFrame and added columns to reference the outcome of the game and goal differential. I split the data with a 70/30 split, resulting in 12,122 training elements and 5,195 testing elements. As discussed below, I implemented different train/test splits and changed how many international games were included in the dataset to find minimal changes in resulting probabilities.

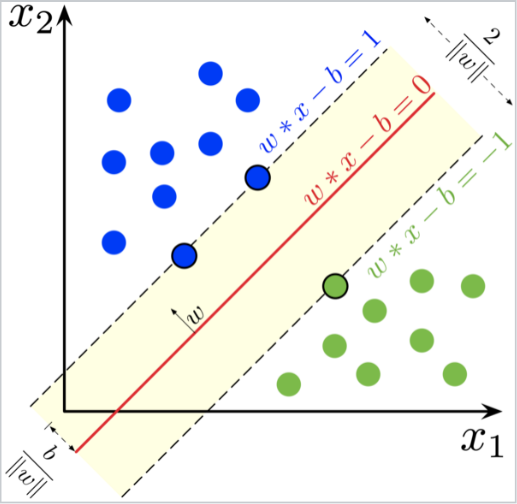
To obtain the datasets I used kaggle for the game results and wrote web scraping code to retrieve the most recent fifa international men's standing. Implementing the code utilized beautiful soup to retrieve HTML information. To parse the HTML code I used 'lxml' and the xpath denoted by the websites in question. I stored the tables in data frames that were then exported to csv files for use in the Jupyter notebook. Using web scraping I was able to access all fifa rankings dating back to 1872. The original dataset including the fifa rankings had data from 1872-2018, to apply this to the 2022 world cup I retrieved the missing data from the last four years.

Methods:

State Vector Machine (SVM)

As mentioned above, state vector machines (SVMs) are used to predict the group stages of the fifa tournament. The SVM is particularly advantageous for this application because it can effectively classify draws as values that fall within the two boundary lines. This effectively creates a disparity between Wins, Losses, and Draws. An SVM accomplishes this classification by finding a hyperplane that has appropriate support vectors that lie on the decision boundary. The support vectors are the closest points to the hyperplane and lie on the decision boundary, meaning they are the edge of the dataset classification, in this case winning, losing, or drawing. SVMs are particularly useful for the group stages because they can classify a game to output as a draw if the passed value is in either edge case, for example if the point falls between the decision boundaries, it is evaluated as a draw for the match. The overall goal of the SVM algorithm is to maximize the distance between the classification points (games) and the hyperplane.

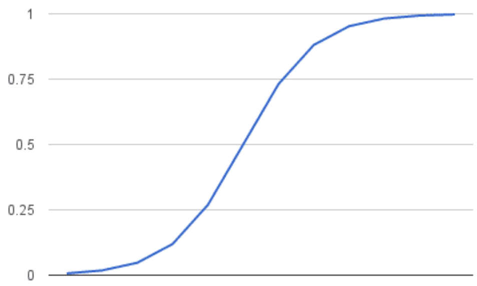
SVMs accuracy are measured by how many of the passed games are correctly classified as wins, losses, or draws. The test accuracy of the SVM was greater than the logistic regression, but that was not the case for training accuracy. Because of this in the experiments, I elected to utilize logistic regression for the knockout stages.



[6] Figure 2. SVM example, depicting the decision boundary 1 and -1, the hyperplane, and support vectors (points lying on the decision boundary)

Logistic Regression

Logistic regression is the go-to method for binary–two value–classification problems. This makes logistic regression extremely useful for the knockout stages when the most important parameter is the team that wins/moves on. Logistic regression gets its name from the logistic function which is used at the core of logistic regression. The sigmoid curve is depicted below in figure 1.



[1] Figure 1. Depiction of Logistic Function modeling the equation

1 / (1 + e^-value)

Below is an example logistic regression equation:

$$y = e^{(b0 + b1*x)} / (1 + e^{(b0 + b1*x)})$$

$$x = 1 / (1 + e^{-value})$$

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data. The actual representation of the model that you would store in memory or in a file are the coefficients in the equation (the beta value or b's) [1]. The binary classification occurs when the probability of a win is greater than or less than a half. Greater than being a win and less than implying a loss.

***Discussion*:**

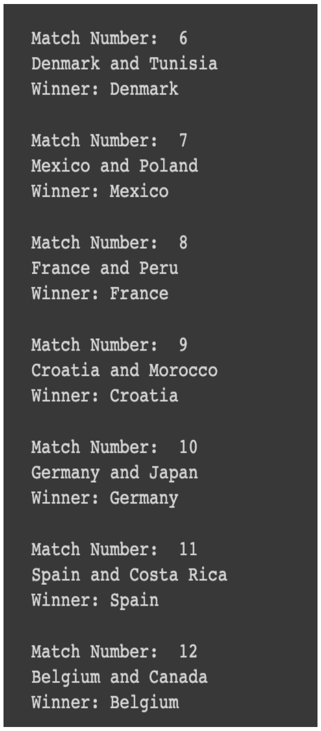
I chose a linear kernel for the SVM since I wanted to perform classification on data points which have a linear decision boundary, so the linear kernel was the most effective kernel that I could choose for the model. I chose C = 1 for the SVM which lets the SVM optimization know how much misclassification I want to avoid for each training sample; I chose a small value of C thus it will tell the optimizer that I want a large margin to separate the hyperplane. I went with the default value of C = 1 since I do not want the learning algorithm to be too sensitive or too less sensitive. For the logistic regression, I imposed a penalty = 12 which causes a penalty to be imposed on the logistic regression model if it has too many variables causing the coefficients which do not contribute much to go towards 0. This is a form of regularization which I think really helps to minimize the adjusted loss function to prevent overfitting or underfitting, which can be seen by the testing and training data being really close in accuracy for the logistic regression. I also use 1000 iterations for the logistic regression which helps in increasing the precision with which the logistic function will fit the data. I split the training:testing data in a ratio of 70:30, meaning that 70% of the data points which will be randomly chosen are being used for training the data, while the other 30% are being used for testing the data. These data points are entries from the results.csv which have a record of all international matches played after 2000. I use cross validation using the cross_validate function from sklearn with scoring=scoring which allows specifying multiple metrics for evaluation and returns a dictionary containing times of fit and score on top of the test score.

Results:

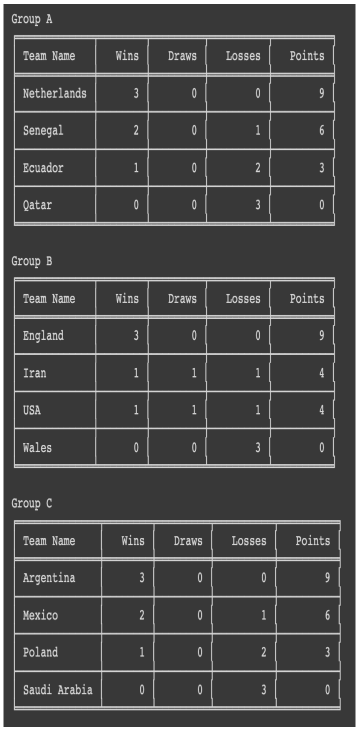
The **SVM** model resulted in an accuracy of 65.2% on the training set, while the testing accuracy was much lower at 55.5%; while the logistic regression model had a training accuracy of 57.6% and its accuracy on the test set was 55.3% which is extremely close to the testing accuracy of 55.5% achieved by the SVM. I also trained a random forest model whose training accuracy was the highest among all the models at 70.4%, however, it had a testing accuracy of only 54.1% which was the worst of all the models I looked at. I decided to proceed with using the SVM and logistic regression models simply because of the fact that they had the highest accuracies.

SVM was used to simulate the group stage matches and the code produced outputs as shown in the figure on the left. Since it is monotonous and difficult to manually keep track of the results of all 48 matches which were played in the group stage, I decided to also output the results in a tabular form which displays how many matches the model predicts each team to win, lose, or draw in their respective groups. The tables also have the number of points the model expects each team to score with 3 points awarded for each win predicted, 1 point awarded for each draw predicted, and 0 points awarded for each loss predicted. This point system exactly mimics the one FIFA uses for the group stages in all it's events including the world cups.

The results obtained and the table created as a result of them are shown below.



```
Match Number:  6
Denmark and Tunisia
Winner: Denmark

Match Number:  7
Mexico and Poland
Winner: Mexico

Match Number:  8
France and Peru
Winner: France

Match Number:  9
Croatia and Morocco
Winner: Croatia

Match Number:  10
Germany and Japan
Winner: Germany

Match Number:  11
Spain and Costa Rica
Winner: Spain

Match Number:  12
Belgium and Canada
Winner: Belgium
```



Group A

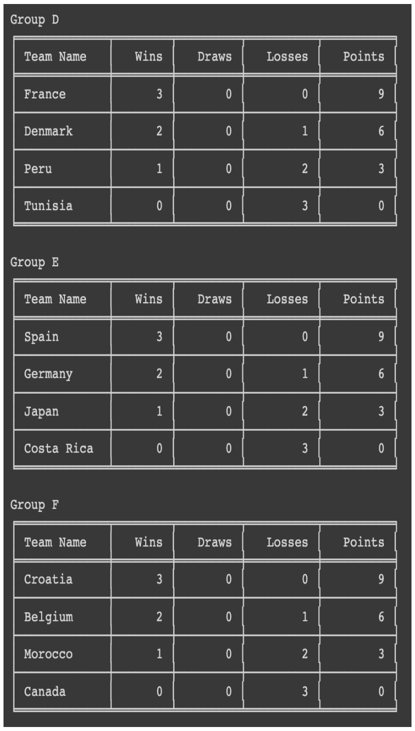
Team Name	Wins	Draws	Losses	Points
Netherlands	3	0	0	9
Senegal	2	0	1	6
Ecuador	1	0	2	3
Qatar	0	0	3	0

Group B

Team Name	Wins	Draws	Losses	Points
England	3	0	0	9
Iran	1	1	1	4
USA	1	1	1	4
Wales	0	0	3	0

Group C

Team Name	Wins	Draws	Losses	Points
Argentina	3	0	0	9
Mexico	2	0	1	6
Poland	1	0	2	3
Saudi Arabia	0	0	3	0



Group D

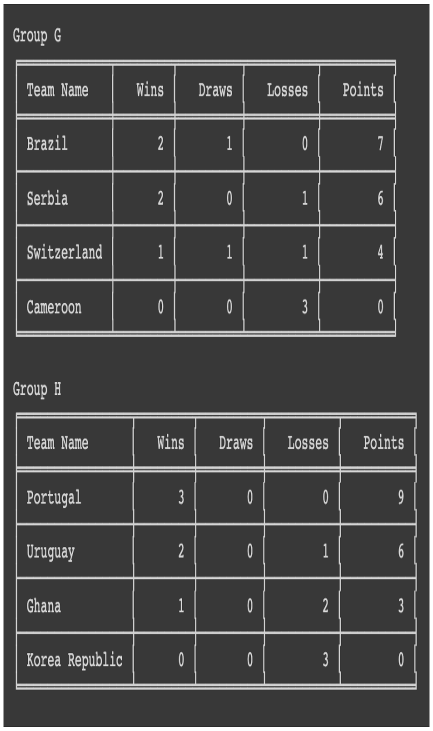
Team Name	Wins	Draws	Losses	Points
France	3	0	0	9
Denmark	2	0	1	6
Peru	1	0	2	3
Tunisia	0	0	3	0

Group E

Team Name	Wins	Draws	Losses	Points
Spain	3	0	0	9
Germany	2	0	1	6
Japan	1	0	2	3
Costa Rica	0	0	3	0

Group F

Team Name	Wins	Draws	Losses	Points
Croatia	3	0	0	9
Belgium	2	0	1	6
Morocco	1	0	2	3
Canada	0	0	3	0

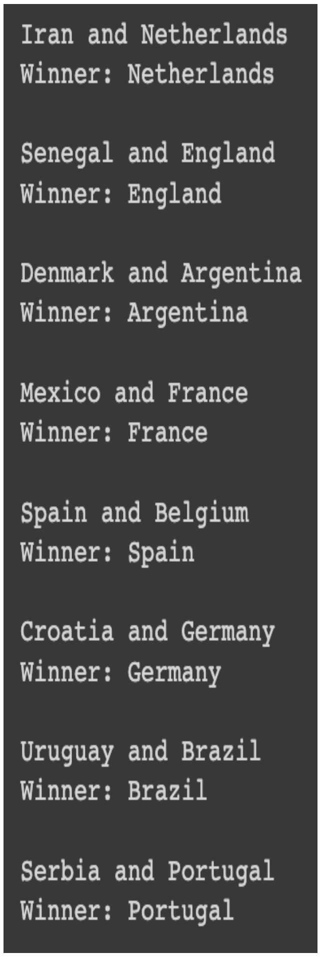


Group G

Team Name	Wins	Draws	Losses	Points
Brazil	2	1	0	7
Serbia	2	0	1	6
Switzerland	1	1	1	4
Cameroon	0	0	3	0

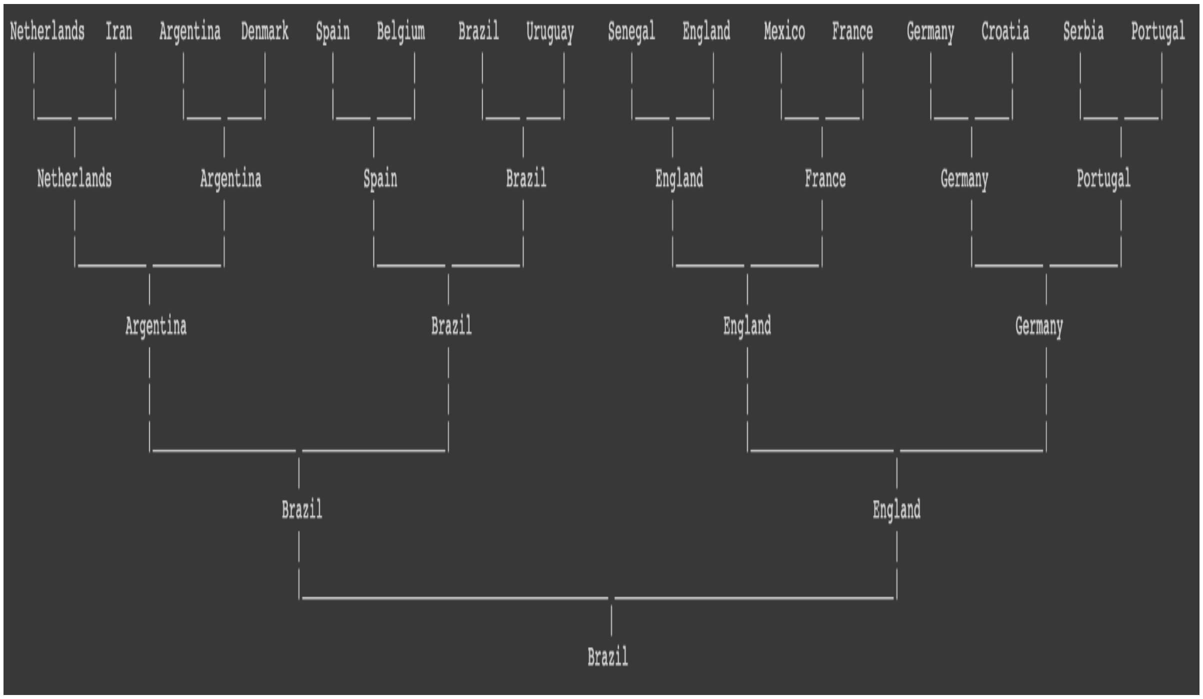
Group H

Team Name	Wins	Draws	Losses	Points
Portugal	3	0	0	9
Uruguay	2	0	1	6
Ghana	1	0	2	3
Korea Republic	0	0	3	0



```
Iran and Netherlands
Winner: Netherlands

Senegal and England
Winner: England

Denmark and Argentina
Winner: Argentina

Mexico and France
Winner: France

Spain and Belgium
Winner: Spain

Croatia and Germany
Winner: Germany

Uruguay and Brazil
Winner: Brazil

Serbia and Portugal
Winner: Portugal
```

Logistic regression was used for the knockout stages and the code outputted the predicted result of each knockout match as shown on the left. The screenshot on the left are results of all the games of the round of 16, and I have output results in a way similar to this screenshot for the quarterfinals, semifinals, and final. Since keeping track of these results is unappealing to do, I decided to also output the results of the knockout stages in a traditional elimination style bracket that I coded and it is shown below.



```
Netherlands  Iran  Argentina  Denmark  Spain  Belgium  Brazil  Uruguay  Senegal  England  Mexico  France  Germany  Croatia  Serbia  Portugal
     |         |       |         |       |       |       |        |        |        |        |       |        |        |        |        |
     |___   ___|       |___   ___|       |___ ___|       |___  ___|        |___  ___|        |___ ___|        |___  ___|        |___  ___|
          |                 |                |               |                  |                 |                |                 |
     Netherlands        Argentina          Spain           Brazil            England            France          Germany           Portugal
          |                 |                |               |                  |                 |                |                 |
          |_______   _______|                |_______ _______|                  |_______   _______|                |_______   _______|
                  |                                  |                                  |                                  |
              Argentina                            Brazil                            England                            Germany
                  |                                  |                                  |                                  |
                  |                                  |                                  |                                  |
                  |_______________   ________________|                                  |_______________   ________________|
                                  |                                                                     |
                                Brazil                                                               England
                                  |                                                                     |
                                  |__________________________________   ________________________________|
                                                                     |
                                                                   Brazil
```

Conclusion:

As can be seen from the above sections, predicting the winner of the FIFA world cup can be tackled as a classification problem with the two classes I used being home team and away team. Predictions regarding which team will win can then be made by picking the team with the highest probability of winning using models such as SVM and logistic regression, both of which are effective for classification problems. I picked SVM (testing accuracy of 55.5%) and logistic regression (testing accuracy of 55.3%) as the classification algorithms due to their testing accuracy being the highest among the models I compared (random forests being the other one) with logistic regression also offering good speed for a slightly poorer accuracy than SVM.

Although the training and testing accuracies that I obtain seem low at 55.5% being the highest achieved with SVM, the results are consistent with most other sport prediction companies whose models also indicate that Brazil are the favorites to win with England and France being tied as second favorites to win. The cause of the low training and testing accuracy lies in the nature of the game of football which is usually extremely unpredictable no matter how much data I collect or which model I use.