ISYE 7406:
Predicting Soccer matches for FIFA World Cup 2022

April 26, 2022

Group 9

Table of Contents

Introduction

Sports betting is a type of gambling that refers to the activity of placing a wager on the outcome of a sporting event. Sports betting has a market size of $219bn with over 30,000 businesses worldwide involved. In other words, predicting the outcome of a sports event is in extremely high demand. The only problem is, to make such a prediction, the human brain is unable to process all the information that is available in this current age. With the World cup starting in about 210 days (as of the time of this document), having a platform or web application that can provide users with the ability to predict game outcomes could hopefully move them closer to winning a sports bet.

The goal of this project is to build a web interface where users can enter two competing teams and a machine learning algorithm would run in the backend to predict the probability of which team would win. The results along with other statistical information would be returned to the user for decision making.

Literature Review

1. Tournament Structure

The FIFA World Cup is an international football competition contested by national teams. This tournament occurs every four years, the current World Cup is taking place this summer in Qatar. The current process involves a qualification phase to determine which teams qualify for the tournament phase. Qualification matches reduce the 200+ teams to 32. Qualification consists of independent regional tournaments: UEFA (Europe) has 13 slots, South America (CONMEBOL) has 4 or 5 slots, Africa (CAF) has 5 slots, Asia (AFC) has 4 or 5 slots, North, Central America and Caribbean (CONCACAF) has 3 or 4 slots and Oceania (OFC) has 0 or 1 slot. The host nation qualifies automatically. Once 32 teams qualify, the group stage beings, followed by the knockout stage.

In the group stage, teams compete within eight groups of four teams each. All teams are seeded. A seed is a competitor or team in a sport or other tournament who is

given a preliminary ranking for the purposes of the draw. Players/teams are "planted" into the bracket in a manner that is typically intended so that the best do not meet until later in the competition. Each group plays a round-robin tournament where each team is scheduled for three matches against other teams in the same group. A total of six matches are played within a group. Considering all the possible outcomes (win, draw, loss) there are 3^6 = 729 combinations possible.

The knockout stage is a single-elimination tournament that begins with a round of 16. Beginning in 2026, FIFA approved a new format, wherein the knockout stage beings with a round of 32 teams.

Six of the eight FIFA champions have won one of their titles while playing in their homeland. The exception here is Brazil who finished as runners-up in 1950 and semifinalists in 2014. England (1966) won its only title while playing as a host nation. Uruguay (1930), Italy (1934), Argentina (1978), and France (1998) won their first titles as host nations but have gone on to win again, while Germany (1974) won their second title on home soil. Many teams have their best performance when serving as hosts. This is something to consider when creating a probabilistic model for soccer predictions.

2. Sports Analytics

The advent of the Internet has created an enormous amount of data, which has facilitated the use of large datasets to identify historical features that can explain statistical significance in many industries. Through the collection and analysis of sports data, sports analytics inform players and coaches to facilitate decision making. As technology has continued to improve, data collection has become much more in depth leading to the development of advanced statistics and machine learning.

Sports analytics has made a significant impact on sports betting/gambling. As mentioned earlier, sports betting has a market size in excess of $219bn. Sports gambling accounts for roughly 13% of the global gambling industry. With this popularity came the development of many sports betting services. Many of these service providers generated billions of dollars in revenue. The ability to create an

accurate predictive model has lucrative potential. It may be surprising to know that a winning percentage on betting above 52% is considered profitable.

3. Prior research on Soccer Predictions

There are numerous methods to evaluate sports and generate outcomes and results. In this paper we will consider the methodology employed by FiveThirtyEight for their soccer predictions. FiveThirtyEight is an American website that focuses on opinion poll analysis, politics, economics, and sports blogging. In 2013 ESPN acquired FiveThirtyEight. Considering the methods employed by various entities may provide us with some insight into how to best create an accurate predictive model.

The predicted forecasts are loosely based on ESPN's Soccer Power Index (SPI), a rating system designed to provide the best possible objective representation of a team's current overall skill level. Every team has an offensive rating that represents the number of goals it would be expected to score against an average team on a neutral field, and a defensive rating that represents the number of goals it would be expected to concede. These ratings, in turn, produce an overall SPI rating.

Due to the low scoring nature of soccer, this sport is especially difficult to predict. To mitigate this variability, three metrics are used to evaluate team performance: adjusted goals, shot-based adjusted goals, and non-shot expected goals. Adjusted goals account for the conditions under which each goal was scored. Shot-based expected goals are an estimate of how many goals a team "should" have scored. Non-shot expected goals are an estimate of how many goals a team "should" have scored based on non-shooting actions they took around the opposing team's goal: passes, interceptions, take-ons and tackles. In addition, nonquantitative variables are considered, such as quality and importance of a specific match. Quality is simply a measure of how good the teams are. Importance is a measure of how much the outcome of the match will change each team's statistical outlook on the season. Match rating is the average of quality and importance.

Looking at this brief explanation of FiveThirtyEight's predictive model, there are numerous variables to consider in the pursuit of an accurate soccer model.

The literature (Demir et al., 2013) mainly focuses on testing the efficiency of the sports betting market. The prediction of game outcomes or comparing the odds of bookmakers by predicted odds and the search for betting strategies which yield significant positive returns have been the core of the market efficiency tests. This study, instead of making any predictions or generating odds to be compared by bookmakers' odds, implements the Fibonacci sequence on draws as a betting rule for 8 European soccer leagues for the seasons from 2005/2006 to 2008/2009. As the odds offered by bookmakers are narrowly distributed, implementing the Fibonacci strategy for 8 soccer leagues of Europe for 4 seasons yields positive return for all cases and controlling with simulated data the strategy is found to be in most circumstances profitable. The results indicate that the bookmakers are inefficient in terms of predicting the draws and the soccer betting markets are inefficient. Therefore, the betters could exploit this inefficiency by following Fibonacci strategy assuming they have enough financial liquidity.

The chapter (Goddard, 2013) evaluates the performance of forecasting models of the efficiency of soccer betting markets. A nontechnical description of a forecasting model is provided, based on a large-scale number-crunching exercise using historical match results and other relevant data. It also examines whether forecasting models of this kind can provide the basis for the development of profitable fixed-odds betting strategies. A prior hypothesis is that opportunities for profitable betting that were available during earlier periods might have been largely eliminated by the increased sophistication of contemporary sports betting markets, greatly enhanced by advances in computing technology.

Why do some soccer bettors lose more money than others? In an efficient prediction market, each gambler should break-even before costs (but losing a constant amount after costs, reflecting the bookmaker's margin). Previous empirical studies across numerous sports betting markets show that bets on longshots tend to lose more than bets on favorites (favorite-longshot bias). The authors (Buhagiar et al., 2018) use 163,992 soccer odds from ten European leagues to test plausible hypotheses around why some soccer bettors lose more money than others. Are soccer bettors with above average losses simply biased, or are their losses driven by betting on events that are inherently unpredictable? They confirm the existence of favorite-longshot bias in soccer in this sample but find another surprising feature of betting on longshots.

Methodology

Overall Concept

1. Calculate teams rating from previous season
    - Calculate teams' market value. We assume the same players from the previous season = Transfermarkt.com.
    - Calculate/Predict number of goals each team is expected to score - Data is available from data set.
    - Calculate/Predict number of goals each team is expected to concede - Data is available from data set.

2. Calculate each team's game performance - Need to find the endpoint for this
    - Number of goals scored
    - Number of shots on target
    - Number of shots off target
    - Ball possession
    - Number of completed passes

3. Using the calculations/predictions from above, we can compare 2 teams in a head-to-head game

Datasets

Data to build the model is collected from multiple sources, including:

1. General Team Statistics

Specifically, we used World Cup team statistics data for 2018, 2014, and 2010; also, the friendly matches statistics from 2020 to 2022. The source of this data is

https://fbref.com. From this dataset, we extracted the following variables to be the prediction features:

- *G-PK*: non penalty goals
- *PI*: number of players
- *MP*: matches played
- *90s*: minutes played divided by 90
- *GLs*: goals
- *Ast*: assistance
- *PK*: penalty kicks made
- *PKatt*: penalty kicks attempted
- *CrdY*: yellow cards
- *CrdR*: red cards
- *G+A*: goals and assists
- *G-PK.1*: goals minus penalty kicks
- *G+A-PK*: goals and assists minus penalty kicks

2. Matches results from 2010 to 2018

The data were obtained in https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017. This dataset provided the following variables to take home team advantage into account:

- *home_team* - the name of the home team
- *away_team* - the name of the away team
- *home_score* - full-time home team score including extra time, not including penalty-shootouts
- *away_score* - full-time away team score including extra time, not including penalty-shootouts
- *tournament* - the name of the tournament

3. Matches results from 2017 to 2022

This dataset (found in http://livescore-api.com/) provided historical data from FIFA World Cup 2018, UEFA, CONMEBOL, CAF, AFC, CONCACAF and OFC matches, and friendly matches.

4. FIFA world ranking

The FIFA world ranking data were obtained in https://www.transfermarkt.com/statistik/weltrangliste.

The information is consistent with the official FIFA website. This dataset provides the updated following national soccer team statistics:

- *Team rank*: Team's FIFA rank
- *Squad size*: Number of players in the team
- *Total value*: Total value of the team in Euros
- *Average age*: Average age of players in the team
- *Points*: Total points of the team

Data Processing

Our goal is to obtain a dataset containing head-to-head statistics of one match in each row.

Steps:

1. We first filter out countries that attended the world cup in 2010, 2014, and 2018 using data obtained from Kaggle [2].
2. Specify a unique team ID for each team.
3. Extract the team statistics from datasets [1], [3] and [4].

During the data preparation, we follow the rules listed below to make sure our data is consistent:

1. Teams' names are checked to be the same in different datasets.

2. Datasets are merged to form one dataset containing matches results from 2010 and average statistics for each team.
3. For the teams' average stats, if there are many records of each team, we calculate the average. And each team's stats are merged with the team results data.
4. A new column was created to show the match result as 1 if Team1 wins, 2 if Team2 wins, and 0 if it is a draw.

We would like to note that only including teams that have ever participated in the world cup since 2010 is due to the data sparsity. The more years we possess, the more missing data exists.

The final dataset for modeling contains 6018 rows and 54columns. Each row represents one match (match would be one of the following tournaments: world cup, friendly matches, all continents' competitions such as UEFA, Copa America, Asian Cup, etc...). Columns are features of both home-team and away-team, such as *home_team_Shots off Goal, home_team_Total Shots, home_team_Blocked Shots, away_team_Shots off Goal, away_team_Total Shots, or away_team_Blocked Shots, etc.*

In the image below, there is a snapshot of the data table:

	home_team.id	home_team.name	home_team_Shots on Goal	home_team_Shots off Goal	home_team_Total Shots	home_team_Blocked Shots	home_team_Shots insidebox	home_tea
0	2	France	4	4	9	1	6	3
1	19	Nigeria	3	5	9	1	4	5
2	2	France	5	8	19	6	12	7
3	6	Brazil	9	7	27	11	16	11
4	10	England	2	6	11	3	6	5
5	2	France	6	0	7	1	2	5
6	4	Russia	5	7	13	1	10	3
7	26	Argentina	3	3	10	4	8	2
8	18	Iceland	6	8	16	2	12	4
9	2	France	5	5	12	2	7	6
10	21	Denmark	5	5	10	0	7	3
11	2	France	4	6	12	2	9	3
12	20	Australia	2	7	14	5	9	5
13	7	Uruguay	3	2	5	0	3	2
14	22	Iran	2	4	7	1	3	4

Modeling

Obtained the statistics for each game, to train the models for competitions between teams, we first aggregate the match data for each team. Since the FIFA World Cup Competition happens every four years, we use all the game data we can obtain between two FIFA years (for example, 2010 and 2014) to predict the results of the next FIFA World Cup Competition (i.e., 2014). Since different teams are from different confederations, the number of games each team plays are usually different from each other. We simply aggregate the match data of the home and away game of each team by taking the mean of the game statistics.

Having the team data aggregated, we then start with the modeling process. For the modeling process, we ignore the seed selection rules of FIFA that the eight seeded teams will not play against each other in the group stage and the grouping rule that no group contains more than two European teams or more than one team from any other confederation, which affects the probability that two teams will play. We also ignore the extra time results in the knockout stage. The model only predicts the full-time match result given two teams meet in a game, ignoring the fact that some teams will never meet in a game.

For training, we first get the aggregated game statistics for the two teams, say Team 1 and Team 2, that have played against each other in a previous FIFA match. Then, we concatenate the two data frames first by the home game data of Team 1 with the away game data of Team 2, then the home game data of Team 2 with the away game data of Team 1 and then drop the team ids to use as two features for training, as home and away game can affect the match result. We label the feature by the full-time result of the FIFA match between Team 1 and Team 2 in that year, which is either 'Win', 'Draw', or 'Lose' or the score, depending on the model. If two teams happen to play twice in a FIFA year, we take the average of the games' full-time result for labeling. We use the data from FIFA 2010, 2014, and 2018 for the modeling. As there are only three years of data, we perform feature selection to avoid overfitting. We use the Preprocess pipeline to perform One-Hot Encoding for the categorical variables and perform PCA for dimensionality reduction. We perform 5-fold cross validation, randomly selecting 80% of the data for training, leaving the rest 20% for testing.

Two models are built from our modeling process. The first one predicts the probability of ‘Win’, ‘Draw’, or ‘Lose’ cases in a full-time match. The second one directly predicts the case of match score, for example, the score case of '1-0', '1-1' or '0-0', etc. We reduced the score difference to a max of 5, which is about 2 standard deviations from the mean. For this task, the model outputs 3 score cases with highest probabilities in each case of ‘Win’, ‘Lose’, or ‘Draw’. For the modeling tasks, we have tried Logistic Regression and XGBoost for classifications. We achieve the best results we can get for each model after hyperparameter tuning. We choose to use the XGboost model for deployment as it gives a slightly higher accuracy, 57%, than Logistic regression.

For prediction, we will have the user input the target team names, say Australia and Germany. We the produce two set of features for prediction by concatenating the data frames first by the home game data of Australia with the away game data of Germany, then the home game data of Germany with the away game data of Australia. After feeding the features into model for each task, we average out the two sets of output as our final prediction results.

Deployment

Our app is hosted on the website: https://share.streamlit.io/lztech-tp/7406datamining/main/web_app.py

The app has two 2 pages which include:

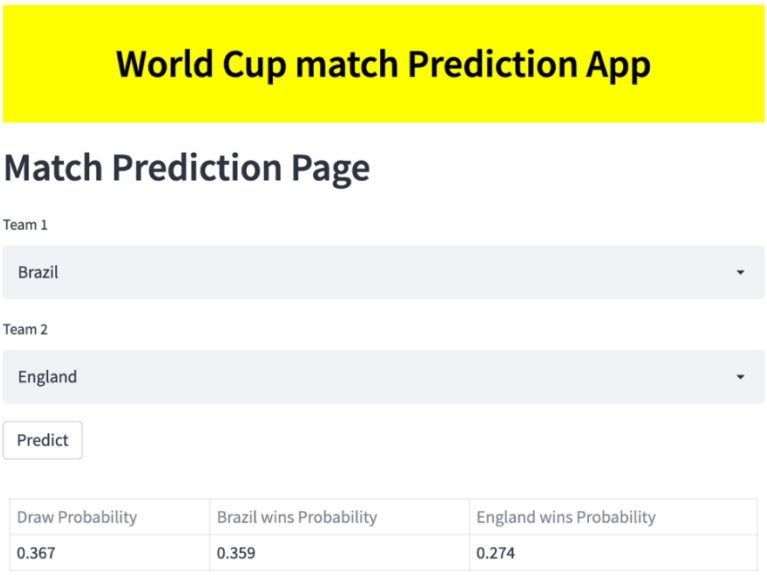
1) The first page has 2 dropdown menus for teams' selection and predict the match winner probability, match score probability, and display the results in tables, and the average statistics of each team.
2) The second page displays the model evaluation metrics as Accuracy, precision, recall, and confusion matrix.

To use the app, the user must select the two teams for which we would like to know the probabilities. Team 1 is the home team and Team 2 is the away team.

The repository the team used can be found in https://github.com/lztech-tp/7406datamining. Streamlit requires the repository to contain the main app file as well as the pkl files and the requirements and docker.

Demonstration

To show how the model works, here we show an example of its use. Brazil is chosen as the home team (team 1) and England as the away team (team 2). Upon clicking predict, we get different probabilities including draw, Brazil wins and England wins.



As seen below, we also show the score predictions depending on the three possible outcomes and the statistics for the two participating teams.

For the example, we see that Brazil is favored to win. This could be because its rank is 2, whereas England is ranked 5. The fact that Brazil is the home team might also influence the winning odds. Also, on average Brazil tends to score more goals, have more assists as compared to England which are some relevant factors that contribute to winning matches.

Match Result prediction

Draw Results

2-2	1-1	0-0
0.369	0.318	0.313

Team 1 win Results

1-0	2-1	2-0
0.399	0.323	0.278

Team 2 win Results

0-2	0-1	1-2
0.368	0.336	0.296

Teams statistics

	Brazil
Team 1 Rank	2.00
Team 1 age	27.70
Total Value in Million €	899.50
Team 1 Points	1823.00
Avg. Minutes played	519.00
Avg. Goals	9.33
Avg. Assists	5.33
Avg. Yellow Cards	3.00
Avg. Red Cards	0.33

	England
Team 2 Rank	5.00
Team 2 age	25.20
Total Value in Million €	1260.00
Team 2 Points	1756.00
Avg. Minutes played	310.50
Avg. Goals	5.00
Avg. Assists	2.83
Avg. Yellow Cards	2.00
Avg. Red Cards	0.00

Our app also allows us to see the model performance. As seen below, our model has an accuracy of around 57%, a precision score of 0.499 and a recall score of 0.57.

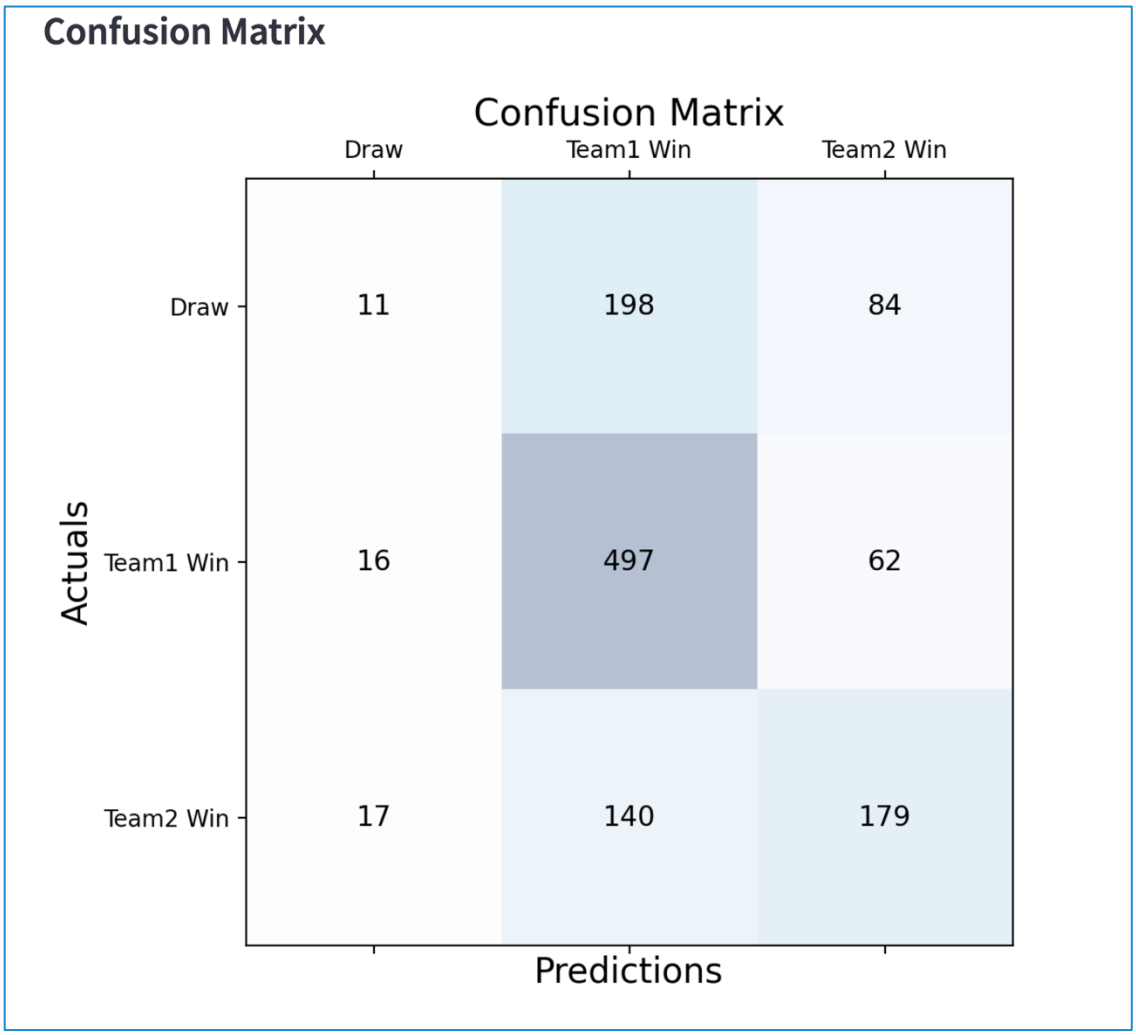
Performance Metrics

```
The accuracy score :0.571

The precision score :0.499

The recall score :0.571
```

The confusion matrix is shown below:



Conclusions

To predict the outcome probabilities of soccer games, we successfully deployed an XGBoost classifier and achieved an accuracy of 57% on the test dataset. We also deployed an app with functionalities of selecting teams and determining the odds of draws and winning.

The project was not without challenges. Some of them are:

- Data availability from endpoints is limited. In fact, many endpoints promise a lot of data, but then the data is sparse for certain years and there are several blank spots.
- Because of the way some endpoints are built, it was necessary to write functions to request the data. For example, we wrote a function that extracted all matches' IDs first so that we could then use those IDs to iterate over other tables and extract data.
- Another limitation we had to overcome is the limited number of requests some endpoints offer to free users. For more requests, it is necessary to upgrade and pay for a subscription. This was solved by inserting pauses in the requests.
- Those endpoints with the highest quality, i.e., the most useful data for the project, are commercial. To use them, it would have been necessary to speak to a salesperson, which was not possible during the time of the development of the project.
- The list of teams that qualified for the World Cup was available in the last stage of the project. For that reason, we spent a lot of time gathering all the data for all countries, which affected the prediction accuracy because of sparse data.
- Streamlit is perfect for rapid prototyping, as we did in the project. However, major challenges exist when building to scale. There is a great need for good platforms that provide flexibility, continuous training, development, and integration, and that are not hard to understand.

References

1. Demir, E., Danis, H., & Rigoni, U. (2013). Is the soccer betting market efficient? A cross-country investigation using the Fibonacci strategy. The Journal of Gambling Business and Economics, 6(2), 29–49. https://doi.org/10.5750/jgbe.v6i2.580

2. Goddard, J. (2013). The efficiency of soccer betting markets. The Oxford Handbook of the Economics of Gambling, 162–171. https://doi.org/10.1093/oxfordhb/9780199797912.013.0009

3. Buhagiar, R., Cortis, D., & Newall, P. W. S. (2018). Why do some soccer bettors lose more money than others? *Journal of Behavioral and Experimental Finance*, *18*, 85–93. https://doi.org/10.1016/j.jbef.2018.01.010