Predicting Sports Bets within the FIFA 2022 World Cup

Aaditya Warrier, Sean Li, Christina Yoh, Brian Janger

Introduction

The FIFA World Cup is the most important international soccer tournament in the world, bringing in over 5 billion projected viewers across the 29 day tournament. It is held every four years and brings together the best national teams from countries around the globe to compete for the title of world champions. Not only does the World Cup provide an opportunity for players to showcase their skills on the biggest stage, it also generates a huge amount of interest and excitement (and thus revenues) and provides people from all over the world an opportunity to come together and celebrate their love of soccer, fostering a sense of global unity and understanding.

For a subset of the followers of the World Cup, sports betting has become a prominent part of the experience. The sports betting industry is a large and growing market that involves people placing bets on the outcome of various sports events. This can include bets on individual games or on the overall results of a season or tournament. According to Bloomberg, a total of $35 billion will be wagered on the 2022 FIFA World Cup, a 65% increase on the previous World Cup.

An integral part of sports betting is the usage of statistics, as it can provide valuable information about the likelihood of certain outcomes in a given game or match. By using statistics, bettors can analyze the true risk on various bets (as opposed to a sportsbook's "odds") and make more informed decisions about which bets to place, giving bettors a better chance of winning and achieving profitable returns.

In short, the goal of our project is to build a predictive model for the 2022 World Cup games using historical international football results and FIFA team rankings in order to provide sports bettors valuable information about the probability of certain outcomes in this year's World Cup. We hope predict goal line bets (point spread/goal differential) as well as 3 way money-line (draw, home win, and away win) bets from the group stage, comparing the odds of respective matches with the odds given by sportsbooks to determine which bets are more likely to be profitable then the odds say.

The dataset we use is a set of thousands of international soccer matches from June 2002 to June 2021, with metrics including team FIFA rank, team FIFA points, match results, offense/defense/midfield score metrics and more. We will first use a Poisson regression model on both home and away team scores to predict score distribution for each team respectively with a lasso penalty to select significant predictor variables and reduce collinearity. Using these relevant predictors, we will then fit a bivariate Poisson model that takes into account the dependency between home and away team goal distributions. Since we have a small number of goals, Poisson regression makes sense as it is intended for response variables that take on small, positive values. Getting results that are probabilistic distributions are important in our case because sports betters are interested in the distribution of results in order to be able to quantify their risk.

Data

Our analysis utilized a few different datasets, which were combined (and later cleaned) into one final dataset. The main dataset was found on GitHub, which gathered data from Wikipedia, the Rec.Sport.Soccer Statistics Foundation (a group which "strives to be the most comprehensive and complete" archive of soccer statistics), and individual soccer team websites. It features the results of 44,341 international soccer matches between 1872 (the year of the first official international match) and 2022.

We also used three other datasets to give us the predictor variables we need to successfully analyze the results and scores of international soccer matches. These included FIFA World Rankings scraped from 2002 onwards, FIFA Player and Team Data, which details the ratings, positions, and other metrics of individual players in the FIFA video game series from the 2015 to 2022 versions of the game, and (box dataset), which tells us additional game specific results like shots on target, possession, red/yellow cards etc.

Data Cleaning

In order for the data pulled from Kaggle to be usable, we had to clean the data first. We first joined our international match dataset from kaggle with the box dataset by year, home team, away team, home team score, and away team score in order to get additional metrics (shots on goal, possession, red/yellow cards etc.) for each match up. In order to make the data more usable, we created new variables that were the differentials between home and away scores for specific categories - for example, we created goalkeeper_differential which equates the home_team_goalkeeper_score - away_team_goalkeeper_score.

Ultimately, we ended up with 13 new variables which are: difference between home and away team goalkeeper score (FIFA game score of highest ranked goalkeeper of team), difference between home and away team defense score (average FIFA game score of 4 highest ranked defensive players of the team), difference between home and away team offense score (average FIFA game score of 3 highest ranked defensive players of the team), difference between home and away team midfield score (average FIFA game score of 4 highest ranked midfield players of the team) , difference between possesion %, difference between shots on target, difference between shots, difference between yellow cards received, difference between shots, difference between red cards received, difference between fouls received, difference between saves made, difference between goals scored, and difference between FIFA rank.

Additionally, we had to create the World Cup 2022 Dataset that has the same columns as our model dataset in order to use it as a final test set. In order to do this, we acquired a CSV with the group stage teams and their FIFA point metrics (offense, defense, midfield, rank, points, goalkeeper), then took averages of the last 5 games as values for other predictors (avg goals, avg shots on target, etc). The reason we had to do this is because typically, in the case of sports betting, the game hasn't occured so those metrics would not be available yet; thus, we are taking the averages as the inputs. Lastly, using these values, we created a dataset that has the same columns as our model dataset by computing differentials as the difference in scores between the home and away team.

Data Cleaning

In order to combine these datasets into a usable one, we first had to clean the data. Upon merging all of the relevant datasets together for the international matches, we discovered that a lot of these matches had non-existent values for box scores or FIFA ratings. We wanted to be able to include all of these potential predictors in our model diagnostics, so we elected to remove these observations from the model. This led to us getting a dataset of mostly recent international matches (as box score data was not widely recorded in the world of soccer until the 2010s).

Our final dataset held data for 786 international matches, including box score data for each team. To make the creation of a predictive model easier, we decided to combine data for each team into a "differential" metric, which found the difference between a statistic for the home team and that same statistic for the away team. Our final set of predictors included FIFA goalkeeper score differential, FIFA defense score differential, FIFA midfield score differential, FIFA offense score differential, percentage of possession differential, shots taken differential, shots on target differential, fouls differential, yellow cards differential, red cards differential, FIFA team ranking differential, whether the match was played at a neutral stadium (i.e. neither team was playing in their home stadium), and the teams playing in the match. For our response variables, we have the home team's score in the match, the away team's score in the match, the computed score differential (home team score minus away team score), and a categorical outcome of the game, where matches were assigned one of three outcomes: the home team winning, the home team losing, and the match ending in a draw (note that for neutral matches, a team is randomly assigned to be the home team).

For our predictor variables above, it is important to note that a positive differential value does not always indicate a good outcome for the home team - for example, a positive shots on target differential indicates that the home team was able to place more shot attempts on the goal than the away team, which is a positive outcome. However, a positive fouls (or yellow and red cards) differential indicates that the home team committed more penalties, which is a negative outcome. This is an important observation to keep in mind when observing our graphs and models featured later in this report.

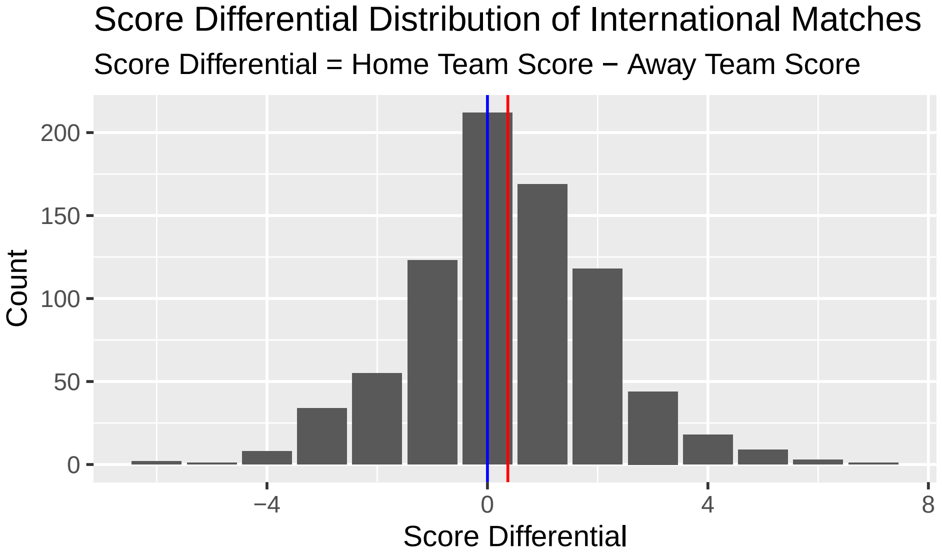
Exploratory Data Analysis

To begin our analysis, we first looked at the distribution of the score differentials and the game outcomes. Since the outcomes of the international matches is a categorical variable that can only take on three values, we show a table approach:

Table 1: International Match Results

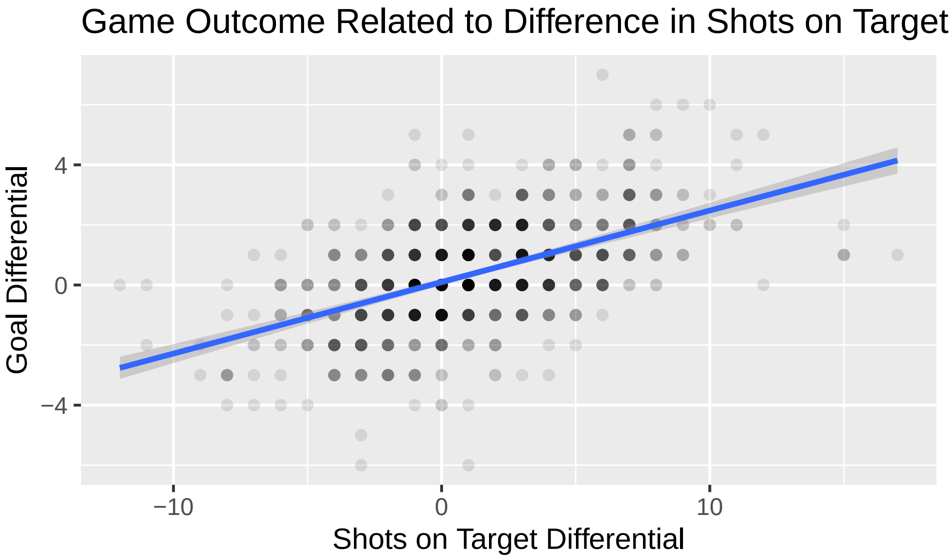
Home Team Result	Count
Draw	201
Lose	227
Win	369

We see that the home team appears to win more often than any other outcome - this will be an important thing to factor into our model as it could accidentally skew outcomes in favor of the home team. Diving further into these games, we see a histogram of the score differentials of all the matches in our dataset:



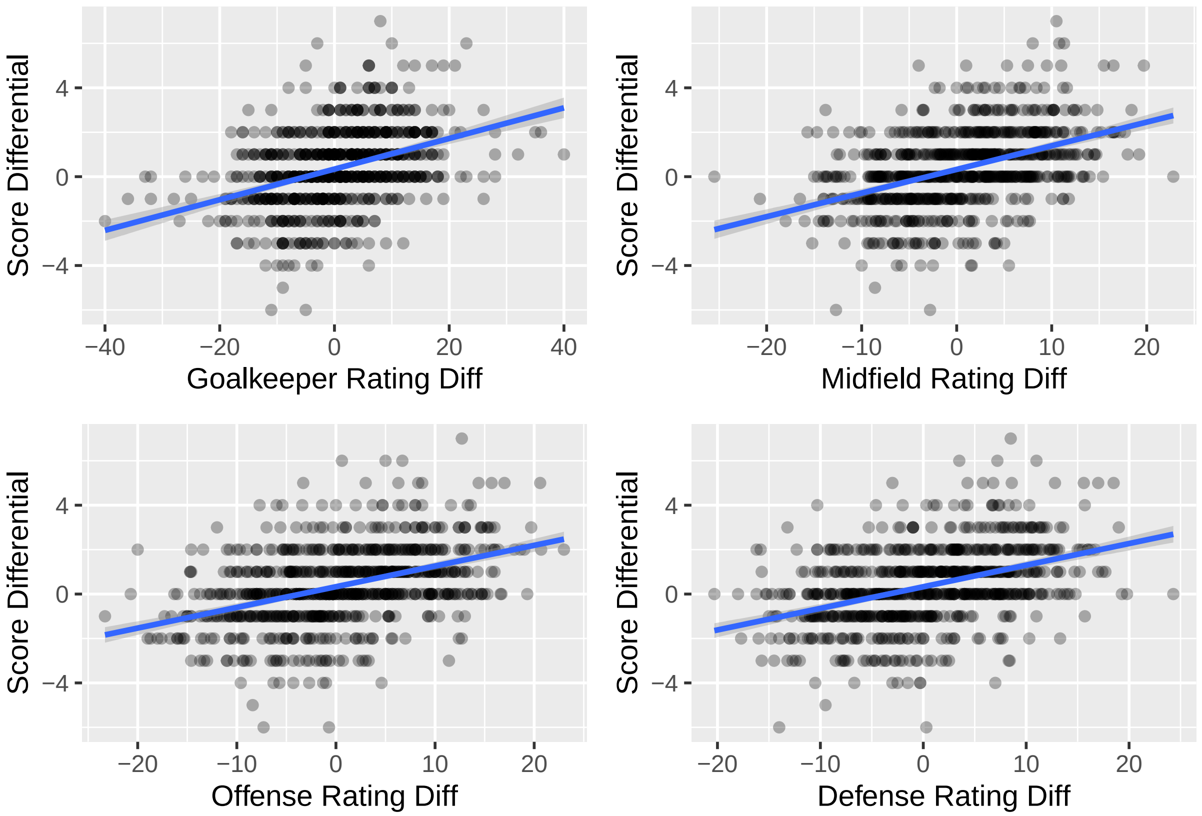
The score differentials range from -6 to 7, and it appears that the distribution of score differentials appears to be close to normally distributed. This graph could indicate that if a team were to win or lose, it is most likely by only a goal or two, as those outcomes make up the bulk of the distribution. The median (blue line) of the score differentials is zero and the mean (red line) of the score differentials is 0.37.

Next, we wanted to see if we could associate any of the predictors visually with the outcome of the game. The predictor variable that appeared to have the largest effect on the goal differential was the shots on target differential:



This made sense to us as an increase in the number of shots on target relative to the other team indicates that there are more opportunities for a team to score. Therefore, an increase in the shots on target differential indicates that they are more likely to win (and win my more goals).

We also saw similar trends (but to a lesser extent) with each of the rating metrics from the FIFA series, shown below:



We see positive relationships with each of these potential predictors and the goal differential. These trends are something interesting we would like to explore in our numerical and probabilisitic models. However, the similarly-distributed scatterplots and lines of best fit could indicate a possible instance of multicollinearity between these predictors - this is something we must explore and do our best to avoid when fitting our models.

Methodology