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 revenue), 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 to 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 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.

Lastly, we also 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 occurred 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.

Exploratory Data Analysis

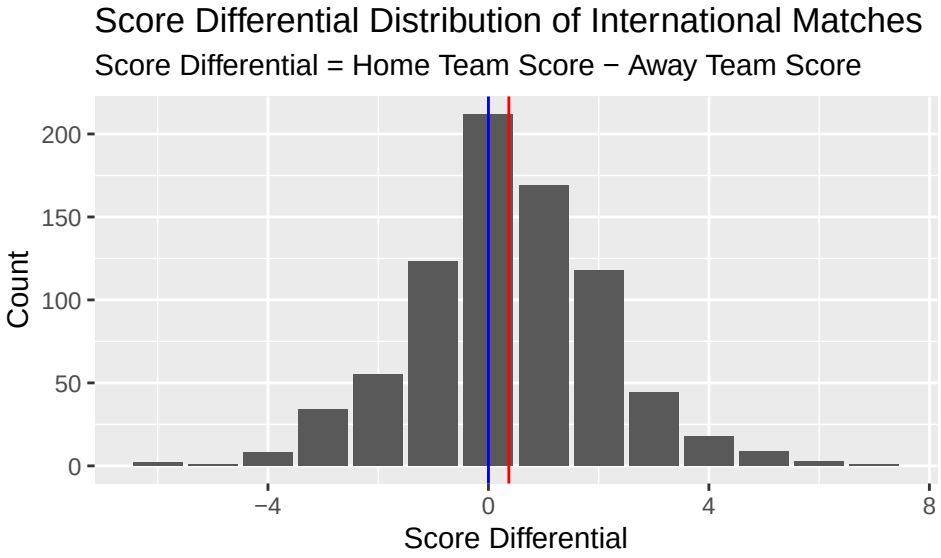
To begin our analysis, we first looked at the distribution of the score differentials and the game outcomes. Since the outcomes of the international matches is a categorical variable that can only take on three values, we show a table approach:

Table 1: International Match Results

Home Team Result	Count
Draw	201
Lose	227

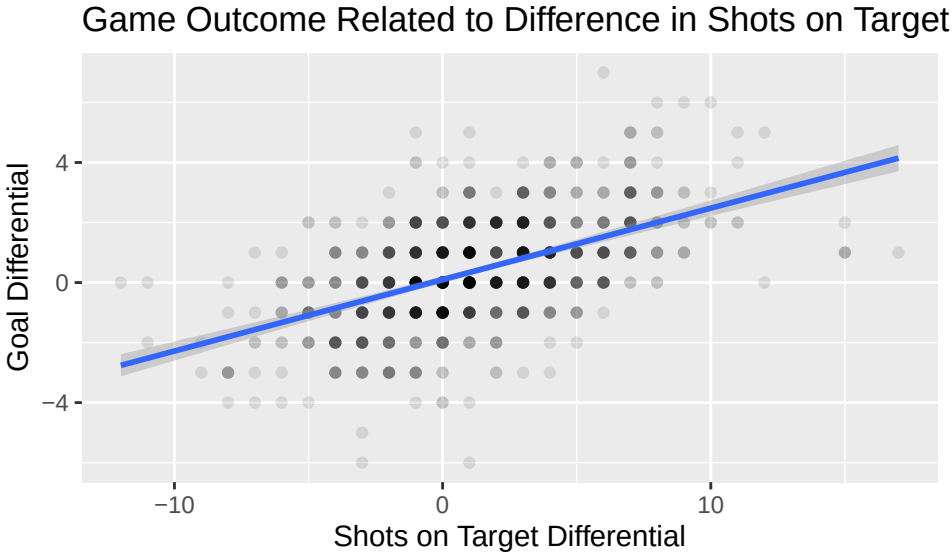
Home Team Result	Count
Win	369

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:



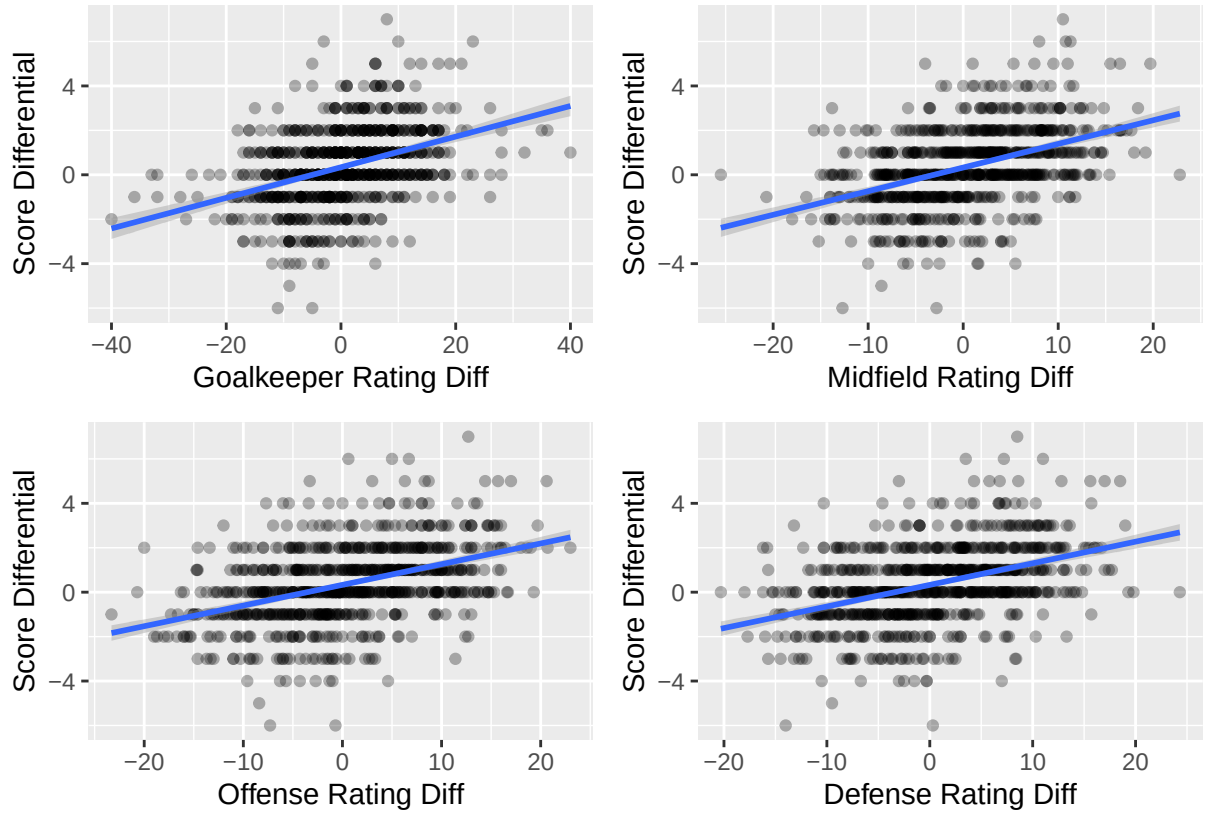
The score differentials range from -6 to 7, and it appears that the distribution of score differentials appears to be close to normally distributed. This graph could indicate that if a team were to win or lose, it is most likely by only a goal or two, as those outcomes make up the bulk of the distribution.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

Modelling Methodology

We initially chose to use a GLM-based Poisson regression model to predict goals for home and away teams in each match, as Poisson regression is most appropriate when the response is a discrete count, as is the case for goals scored in a match. We wanted to perform both regularization and variable selection, as our dataset had a very large number of correlated predictors and we suspected that performance could be predicted using a much sparser set of variables; therefore, we introduced a LASSO penalty using the glmnet package, and 5-fold cross validated for the optimal value of lambda. However, we quickly realized that since the outcome of a soccer game is very dependent and how two specific teams interact with each other, we could not use univariate models that would predict a team's score in a vacuum. Hence, we finally decided to use a bivariate poisson regression model to account for these possible correlations.

The general form of this model is as follows (cite later):

Consider random variables X_k, K = 1, 2, 3 which follow independent Poisson distributions with parameters λ_k, respectively. Then random variables X and Y are given by $X = X_1 + X_3$ and $Y = X_2 + X_3$ and jointly

follow a bivariate Poisson distribution. So, E(X) = $\lambda_1 + \lambda_3$ and E(Y) = $\lambda_2 + \lambda_3$. In our case, X represents goals scored by the home team, while Y represents goals score by the away team.

When we add covariates, we then have:

$$(X_i, Y_i) \sim BP(\lambda_{1i}, \lambda_{2i}, \lambda_{3i})$$
$$log(\lambda_{1i}) = w_{1i}^T \beta_1$$
$$log(\lambda_{2i}) = w_{2i}^T \beta_2$$
$$log(\lambda_{3i}) = w_{3i}^T \beta_3$$

where i = 1, . . . , n, is the observation number, w_i is a vector of predictors for the i-th observation used to model λ_{ki}, and β_k denotes the corresponding vector of regression coefficients K = 1, 2, 3.

Upon research, we discovered that bivariate models with LASSO regularization have not been invented as of yet (or at least not in a form that is easy to implement). To simulate the potential benefits of this framework as much as possible, we decided to screen our variables using the results of our univariate LASSO models, choosing the sparser set of variables among the home and away models as final predictors in the bivariate model.

As sports bettors are generally interested in a spread/probabilistic outlook on game outcomes to make an informed decision, we decided to run 100,000 Monte Carlo simulations of our games to obtain estimated odds/probabilities of each game closing with a home team win, loss or draw. We did this by drawing randomly from bivariate Poisson distributions, whose parameters were outputted/estimated by our regression model based on predictor values.

During this process, we tried a number of other models and datasets to build the most robust model - this included a multinomial model that predicted outcomes, as well as a bivariate Poisson model on a far larger dataset that contained possession stats, shot accuracy etc. for individual matches, which we could use to calculate running averages for each team. However, when comparing accuracy using misclassfication error from our Monte Carlo draws, using 5-fold cross validation on our training set as well as predicting outcomes for the test set (i.e 2022 games), we found that our original dataset and model worked best and were the most statistically robust.

Finally, we

Assumptions:

Clearly, independence cannot be satisfied here as each team's result in a match is not independent of their result in another; however, for the purposes of our analysis, we can assume that our fairly large sample size can counter this to some extent and stay. Caveating this, the two primary assumptions that need to be satisfied are:

1. Ensure that the response is a count, and that these counts are Poisson distributed. This means mean should be roughly equal to variance for goals scored for each team historically.
2. There are linear relationships between log(response) and changes in predictor variables. Details are in the appendix.

The first of these is undoubtedly the most important - it is self-evident that the response is indeed a count, but

Model Results

Interpretation

Our final predictors for our bivariate poisson model were defense_differential, midfield_differential, and offense_differential. These predictors suggest that the FIFA rankings of the players on the respective teams

are the most indicative in predicting match score outcomes. For every one unit the home defense is better than the away team, we can expect the rate parameter for home goals to be multiplying by a factor of $e^{0.0196-0.2269842040} = 0.812$. For every one unit the home midfield is better than the away team, we can expect the rate parameter for home goals to be multiplying by a factor of $e^{0.0150079479+0.1251729042} = 1.15$. For every one unit the home offense is better than the away team, we can expect the rate parameter for home goals to be multiplying by a factor of $e^{0.0101+0.0692} = 2.017$.

For every one unit the home defense is better than the away team, we can expect the rate parameter for away goals to be multiplied by a factor of $e^{-0.0009854082-0.2269842040} = 0.796$. For every one unit the home midfield is better than the away team, we can expect the rate parameter for away goals to be multiplied by a factor of $e^{-0.02811+0.125} = 1.101$. For every one unit the home offense is better than the away team, we can expect the rate parameter for away goals to be multiplied by a factor of $e^{-0.02289+0.0692} = 1.047$.

	x
(l1):(Intercept)	0.1717589
(l1):defense_differential	0.0237813
(l1):midfield_differential	0.0095639
(l1):offense_differential	0.0133452
(l2):(Intercept)	0.0591298
(l2):defense_differential	-0.0044015
(l2):midfield_differential	-0.0250488
(l2):offense_differential	-0.0237410
(l3):(Intercept)	-2.9651838
(l3):offense_differential	-0.0253598
(l3):defense_differential	-0.0890923
(l3):midfield_differential	0.1558049

From the model coefficients, it's interesting to note that some of the conclusions for the home and away team goals are seemingly contradictory - for example, our model predicted that for every one unit the home midfield is better than the away team, we can expect the rate parameter for home goals to be multiplying by a factor of $e^{0.0150079479+0.1251729042} = 1.15$, but also predicted that for every one unit the home midfield is better than the away team, we can expect the rate parameter for away goals to be multiplied by a factor of $e^{-0.02811+0.125}$=1.101. However, since 1.15 is greater than 1.101, we can say that the goal differential (home-away) will still be positive - in the case that home midfield is one unit better than the away team, the rate parameter for home goals is greater than the rate parameter for away goals. The coefficients for the defense seem a little bit counterintuitive, as it states that as defense score increases, the rate parameter for home goals decreases, implying that teams should weaken their defense to score more, which doesn't really make sense. However, a possible interpretation is that we can look at a focus on defense, midfield, and offense as a spectrum – the more a team focuses on defense, to less they focus on offense, and thus may be less likely to score.

For sports betters, the main takeaway is that a team's player composition - their defense, midfield, and offense player rankings - are the most important metrics to investigate (over other metrics like shots or saves in past games) when the results of a matchup. In addition, the composition of a team and their respective strengths in defense, midfield, and offense, is also indicative of how many goals they end up scoring compared to the other team and thus the final outcome of the game.

Estimated Probabilities

Predictive Accuracy

Based on our model's probabilities, we were wanted to see how often our model's most likely outcome for a given match would be the correct one. We calculated our confusion matrix below:

Table 3: Confusion Matrix for Model

	Likely Loss	Likely Win
Draw	2	8
Loss	12	7
Win	5	14

This implied a win accuracy of about 74%, a loss accuracy of about 63%, and an overall accuracy of about 54%. Note that our model did not predict any draws at all, despite there being 10 draws across 48 games, which hence put our draw accuracy at zero. Although this last finding seems problematic, it must be acknowledged that draws are notoriously difficult to predict accurately - in fact, using an aggregation of sports books odds from OddsPortal (a popular sports betting website) to predict outcomes (cite) leads to exactly the same result:

Table 4: Confusion Matrix for Sports Books

	Likely Loss	Likely Win
Draw	3	7
Loss	10	9
Win	5	14

From the ultimate effect and implications of this will be discussed briefly in our conclusion.

Conclusion

To summarize our work - our model tackles the most important bet in soccer, the 3-way moneyline (win, loss, draw). For the group stage in the 2022 World Cup, we came up with relative probabilities for each possible match outcome based on modeling the number of home goals and away goals on a bivariate Poisson model. We further analyzed our model coefficients and what they said about drivers of performance in soccer, as well as their implications for sports bettors. Finally, we compared our model's average accuracy over all matches to the accuracy implied by sports books odds.

In closing out our analysis, we wanted to see if our model could actually make us money on average; with payouts based on the pregame moneyline odds taken from an aggregation of sports books, we chose to simply bet on the most likely outcome for each match based on our model. Of course, this is not always the most optimal betting strategy - importantly, the objective of the model is to mark odds as accurately as possible for each match, not tell bettors who to bet on - but it would give us an idea of how profitable the model would be if we were to make the safest bets possible.

For moneyline bets, our model would have obtained a net profit of -$327, i.e it would lose 327 dollars over all 48 bets across group stage matches. This is assuming one placed a flat $100 bet for each match on the most likely line as predicted by our model, as previously mentioned. However, while our model lost money, it still **outperforms** a bettor who places the same 100 dollars on the most likely line as indicated by the sports books' odds, who would lose roughly $2600. This makes perfect sense - according to most sports betting experts, the optimal betting strategy is to bet on underdogs at least some of the time i.e bet on less likely outcomes in at least some matches. This is because lines for unlikely outcomes have far greater payout, so while betting on likely outcomes would probably end up in making the right call more often, this does not ensure long-term profit. However, it is reassuring all the same that our model's had odds and calls marked more accurately for these "safe" bets.

Most of our work's limitations stem from the nature of the problem we are trying to solve - sports are inherently hard to predict, which is why any model we built did not have great accuracy. If building high performing models was easy, then sports books would simply go out of business! Much of the intrigue of

sports betting is indeed knowing when to make the call on the underdog, something our model simply cannot capture. Furthermore, it is difficult to take into account the interaction between two specific teams, as they prepare customized game plans that differ based upon opponent that are not released to the public. These intangibles were not accounted for by the predictors for the models were built. Sports bettors should heed also caution when using our model (or any model) to accurately predict draws; as we saw before, it predicted none of the 10 draws across 48 matches correctly. This limitation is likely linked to the low-scoring nature of soccer. Future work may include building models to target other bets, such as over/under on total goals scored or various team/player prop bets.