Predicting FIFA 2022 World Cup Results *

Luis Simplicio Ribeiro Elie Attias Isidora Diaz

December 12, 2022

1 Motivation & Context

The FIFA World Cup is an international soccer football tournament played by men's national teams every 4 years. The 2022 version is the 22nd World Cup, taking place in Qatar, starting in November 20th until December 18th. Traditionally the tournament has had 32 participating teams, and constitutes one of the most viewed sporting events in the world.

The tournament is organized in 6 stages: Group, Round of 16, Quarter-finals, Semi-finals, Third-place play off and Final. The Group stage is contested in eight round-robin groups, meaning that every team meets every other group participant in turns. In contrast, all the next stages work in a elimination fashion, where the selected teams from the Group stage play in at least one additional encounter round, and in each subsequent stage the winner competitors from each encounter progress to the next one. As the teams advance, the number of competitors and encounters decreases, until only two teams face each other in the Final.

The previous World Cup was held in Russia in 2018, where the French team beat Croatia in a match 4-2 match. In this project we will leverage various sources of team and player data to construct regression and classification models to predict matches' outcomes for the FIFA World Cup 2022. We will evaluate those models against the baseline predictions using only the FIFA Ranking measure of team strength.

2 Data Description

We worked with two data sets: FIFA World Cup and FIFA 22 Complete Player Dataset, performing different custom pre-processing steps that allowed us to perform predictions at the team level.

- **Fifa World Cup**
  This data set contains historical results of the World Cups starting from 1930 until 2014 from FIFA's World Cup Archive. It contains 3 related datasets: Matches, Players, and Cup. Matches is our dataset of interest, because it contains detailed information about opponent teams (encoded using FIFA Alpha-3 codes), number of goals and World Cup stages. The dataset originally had 4572 data points, with 20 features each. But, 3722 of these data points contained NaN values, for all the features. All these rows with NaN values were dropped, remaining 850 rows in total.

  The following features were removed since they are not very informative about the result of the match: `Datetime`, `Stadium`, `City`, `Attendance`, `Half-time Home Goals`, `Half-time Away Goals`, `Referee`, `Assistant 1`, `Assistant 2`, `RoundID`, `MatchID`.

  Some other features were used to infer relevant information. For instance, the data only directly tells us the number of goals scored by each team, but not who was the winner (if we have one). Because of this we constructed a variable to indicate if the match resulted in a win for the Home Team, Away Team, or a tie based on the goals' difference for a given match. Nevertheless, ties are only allowed during the Group stage (because of the round-robin structure), this means that for the rest of the stages we needed to use an additional feature to indicate the result of the game when two teams where on a tie after a 90 minute play. For this purpose we used the `Win conditions` variable, that described with words the special actions taken to resolve ties during the elimination stages. The content of these descriptions usually had this format: "Team X wins after extra time" or "Team Y wins on penalties (4 - 3)".

*`https://code.harvard.edu/ela324/Project_209_FIFA/`

We also perform one-hot encoding to the categorical variables: `Stage`, `Home Team Initials` and `Away Team Initials`. After this, we end up with a matrix with 850 rows and 174 columns.

#	Column	Non-Null Count	Dtype
0	Year	852 non-null	float64
1	Datetime	852 non-null	object
2	Stage	852 non-null	object
3	Stadium	852 non-null	object
4	City	852 non-null	object
5	Home Team Name	852 non-null	object
6	Home Team Goals	852 non-null	float64
7	Away Team Goals	852 non-null	float64
8	Away Team Name	852 non-null	object
9	Win conditions	852 non-null	object
10	Attendance	850 non-null	float64
11	Half-time Home Goals	852 non-null	float64
12	Half-time Away Goals	852 non-null	float64
13	Referee	852 non-null	object
14	Assistant 1	852 non-null	object
15	Assistant 2	852 non-null	object
16	RoundID	852 non-null	float64
17	MatchID	852 non-null	float64
18	Home Team Initials	852 non-null	object
19	Away Team Initials	852 non-null	object

Table 1: World Cup Matches dataset overview

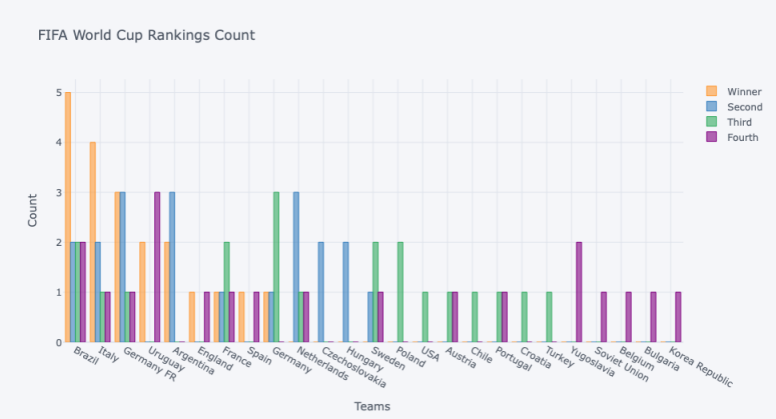

Figure 1: FIFA World Cup Rankings Count

- **FIFA 22 complete player dataset**

This data set contains player's information scraped from the latest edition of FIFA videogame from FIFA 15 to FIFA 22 (Career Mode), a football simulation videogame. We believe that FIFA 22 should be a good proxy to capture each team's current overall performance at the player level, that we will later aggregate to feed into the Poisson regression at the team level to predict number of goals in a match. It has +19k data points (N=19.239) of active football players around the world and 110 features (P=110).

As it is possible to see in figure 2, even though the data stems from a video game, overall distribution of top players between countries fits our prior guess, with countries from Latin America and Europe showing the best performance. Figure 2 additionally explores overall rating distribution for countries that have won a World Cup before.

We performed a thorough data cleaning and feature selection process with the objective to be able to identify what are the most relevant features to predict overall player performance.

First, we needed to standardize the countries' names (originally in long name string) into Alpha-3 ISO codes and FIFA team member codes, that will later allow us to merge this data with the FIFA World Cup Matches.

Second, we constructed 3 main feature categories: to-drop, numerical and categorical. The to-drop list contains mostly columns with information that is not relevant for our target dependent variable (`overall`), including player's game id, league name, jersey number, url, to name a few. We differentiate between numerical and categorical variables because they required different pre-processing to be able to be used in a regression model. For numerical features we imputed the variables with missing values (p=10) with the mean conditional on the player's position. For categorical features we imputed variables with missing values (p=3) with the mode conditional on the player's position. Using a conditional mean and mode is important because usually players are specialized in their respective positions, which means that their ability set be very different depending on where they are positioned in the field: for example, goalkeepers are in average consistently slower than midfielders, so if we where to impute a goalkeeper's speed with the unconditional mean, we would be artificially assigning a higher value than is actually expected given that a player is a goalkeeper.

- **International football matches since 1872**
  This dataset includes 44,152 results of international football matches starting from the very first official match in 1872 up to 2022. It contains the following features `home score`, `away score`, `country score`, `tournament` ,`date` among many others. In other to make predictions based on modern events, we only considered matches that took place after the 2018 World Cup.

- **FIFA World Ranking 1992-2022**. This data set consists in the FIFA ranking evolution of countries since 1992.

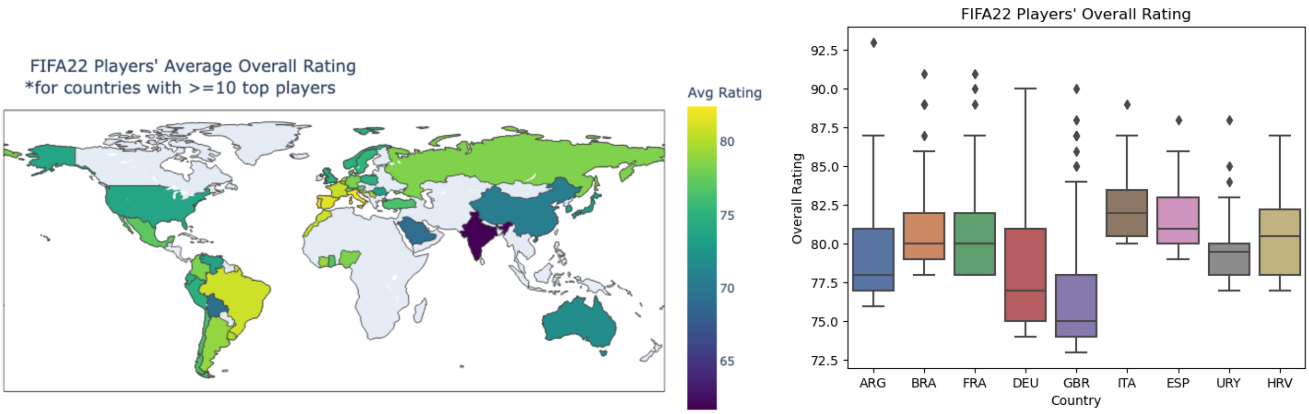

Figure 2: (left)Average overall rating for countries, (right)Players' overall rating distribution

3 Project Question

How accurately can we predict FIFA World Cup 2022 results using regression and classification models at the team level?

4 Models

4.1 Cross Validated Lasso Regression for feature selection

Lasso (least absolute shrinkage and selection operator) regression is a type of regularized regression where the magnitude of the coefficients and the magnitude of the error is penalized through a L1 penalty. Lasso regression allows to reduce overfitting by reducing model complexity, by shrinking the features' coefficients. In Lasso regression

this penalty can result in the elimination of some variables when their coefficients are shrunk to zero, which is why it is a suitable model for feature selection.

$$\min_{\beta_0,\beta} \left\{ \frac{1}{N} ||y - \beta_0 - X\beta||_2 \right\} s.t. ||\beta||_1 \leq \lambda \tag{1}$$

Because the penalty strength is a hyper parameter, we made use of cross validation to select the most appropriate value for λ

4.2 Poisson Regression to predict the number of goals of each team

Poisson regressions are normally used to predict data counts. They assume that the response variable has a Poisson distribution and that the logarithm of its expected value can be modeled by a linear combination of parameters to be estimated.

We want to predict how many goals each team will score at each match, from where we will later determine a winner based on the team with most predicted goals.

To do so, we define the response variable Y_j^i as: the number of goals team A scored in the i^{th} match, for $i \in \{1, \cdots, m\}$. We assume this random variable follows a Poisson distribution : $Y_j^i \sim Poi(\lambda_j^i)$ where λ_j^i is a parameter we wish to estimate from our data by analyzing the number of goals teams scored and conceded in the past.

Our design matrix is denoted as $\mathbf{X}$ where each predictor X_j is composed by a set of m features $X_j = (X_j^1, \cdots, X_j^m)$,, that could include the opponent team, Home or Away condition, the Stage, and aggregated players' features.

In order to relate the predictors $\{X_j^i\}$ to our response variable Y_j^i we will use properties of generalized linear models. Generalized linear models, to which Poisson regression belongs, are more flexible because they allow the response variable to be non-normally distributed.

$$E(Y_j^i) = \mu_j^i = g^{-1}\left(\beta_0 + \sum_{k=1}^{m} \beta_k X_{ik}\right)$$

The assumptions for Poisson regression are the following: (1) Y-values are counts, (2) Counts must be positive integers, (3) Counts follow a Poisson distribution, (4) Features are continuous, dichotomous or ordinal, (5) Observations are independent.

We performed a visual inspection of our dependent variable (number of goals per match) to confirm if the matches dataset distributed close enough to a Poisson distribution. Figure 3 shows that a priori it is safe to affirm that the goals in the world cup matches follow a Poisson distribution.

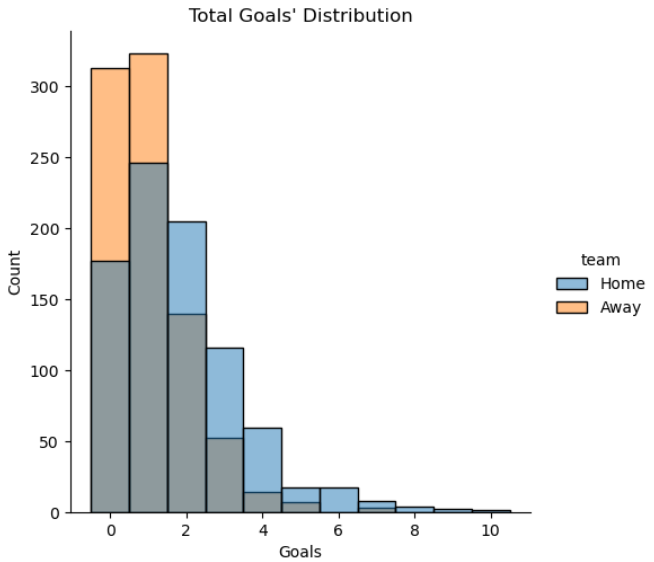

Figure 3: Goals distribution, FIFA Matches dataset

4.3 Logistic Regression to predict final outcome of a game

Another alternative to accomplish our objective is to use a classification model. In this case, we are not interested in the number of goals each team scores during games, but rather interested in the final outcome of a game. We aim to predict if a team i will win, draw, or lose a game against team j. We can assign to each outcome a one-hot encoded variable

$$Y_{ij} = \begin{cases} 1 & \text{if team } i \text{ wins against team } j \\ 0 & \text{in case of draw in group phase .} \\ -1 & \text{if team } i \text{ loses against team } j \end{cases}$$

We build a logistic regression model where the outcome of the match facing team i against team j would have probability: $P(Y_{ij}) = \frac{1}{1+e^{f(\mathbf{X},i,j)}}$ where $f(\mathbf{X},i,j) = \sum_{t=0}^{T} \beta_t X_{ij}(t) + \sum_{t=0}^{T} \alpha_t h(t)$ where $X_{ij}(t)$ is the FIFA ranking difference between team i and team j at time t, and $h(t)$ determines if the game is friendly or not.

5 Results

5.1 Feature Selection

Because of the high number of available features, we used a Cross Validated Lasso Regression selection model to predict overall rating performance and be able to select only the most relevant features.

As we know, in regularized regression such as Lasso, it is necessary to standardize the variables before fitting the model in order to avoid an unfair penalty due only to the features' magnitude difference. For this reason, we standardized all the variables after imputation and encoding to have mean 0 and 1 standard deviation. After fitting LassoCV and imposing an additional 0.5 magnitude threshold over the coefficients, we were able to reduce our features of interest from 111 to 21: `value_eur` (player's monetary value in euros), `pace`, `shooting`, `passing`, `dribbling`, `finishing`, `accuracy`, `attacking` (average between finishing and short-passing), `skill` (average between dribbling and ball control), `movement` (average between acceleration, sprint speed and reactions), `mentality` (average between interception, positioning, vision and composure), and `goalkeeping` (average between all goalkeeping abilities).

After this process, we aggregated these results at the team level and used them as features in the Poisson regression models (see next subsection).

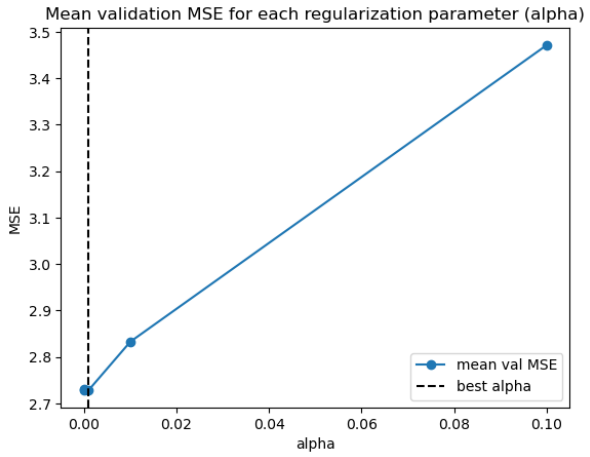

Figure 4: Cross validated Lasso Regression

5.2 Poisson Regression

In order to perform a Poisson Regression model we unstuck the pair structure of the matches in the dataset into a long format, where each row has a single team and its goals in a match. The dependent variable in this case

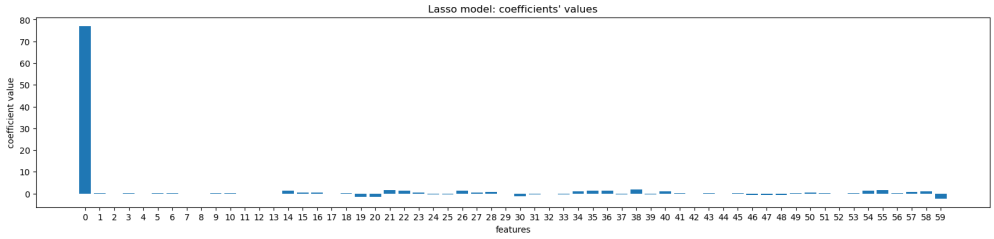

Figure 5: Cross validated Lasso Regression Coefficients

D^2 score	Train (%)	Validation (%)
Poisson, Baseline	24.8	24.1
Poisson incl. Stage	25.5	26.1
Poisson incl. Players	25.0	13.1

Table 2: Poisson Models' scores (each building on top of each other)

are the number of goals, while the independent features for the baseline model are the team's country, the team's opponent team, and a dummy variable to signal if the team was playing as Home or Away.

This baseline model has a percentage of deviance explained (D^2) of 24.8% in the train set and 24.1% in the validation set.

After we incorporate the stage at which the match was played, we obtain only a slightly higher D^2, with 25.5% for the train and 26.1% for the validation set.

We also merged the matches dataset with the players' dataset (taking the mean for each team for the features that were deemed as relevant in the previous section), however the deviance explained in the test set dropped more than 10 percentage points, obtaining 25% in the train set and 13.7% in the validation set.

Given the low percentage of deviance explained by these models, we decided not to move forward to predict wins and defeats. We could also conclude that the players dataset wouldn't be useful for prediction. However, even though the accuracy of the Poisson Regression is lower than we expected for prediction, we decided to explore the data using predicted probabilities to see if there were clear differences for the teams that made it to the Quarter-final stage. As we can see in figure 6, most of the teams have very similar prediction curves for different number of goals, with Morocco being the only team that has a high probability to not score any goals (0 goals), of near 50%. This fact makes it even more surprising that the Morocco team managed to win against Portugal this past Saturday and make it to the Semi-final stage.

In figures 7 and 8 we can see the simulated goal difference for 2 selected matches that happened during the Quarter-final stage. We achieve these plots using the Skellam distribution that results from taking the difference between two Poisson distributions. We can observe that in figure 7 the distribution is skewed to the left, meaning that in our simulation Brazil had slightly better probability to beat Croatia by a 1 goal difference. For the match between France vs. England we would expect the match to be very tight, as we can see that the goal difference distribution is centered around zero for this pair.

5.3 Classification Regression

A more successful approach was to use different classification models to predict if a match ended in a victory, defeat or tie, for a specific team. In our first try, we used the Matches Dataset using paired information, in other words, the features of both teams were stacked together, to predict if the home team would win, lose, or if that particular match would end as a tie. After preprocessing the dataset, we basically used categorical features, such as who is the home team, and in which stage the teams are playing, to train our model. In order to capture the time dependence on the performance of each team, such as, average of goals in the last world cups, we tried to use a feature which calculates a weighted average of the average number of goals scored and conceded by the team in previous participations on the world cup:

- $\text{gs}_A^j \propto \sum_{i=1}^{j-1} \frac{1}{\gamma^{m_i}} \times$ (average number of goals scored by team A in its i^{th} participation)

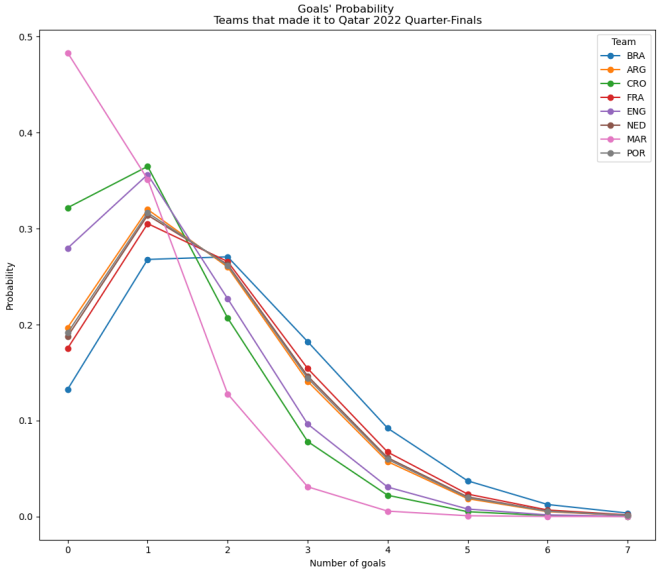

Figure 6: Predicted goals for teams in Quarter-finals 2022

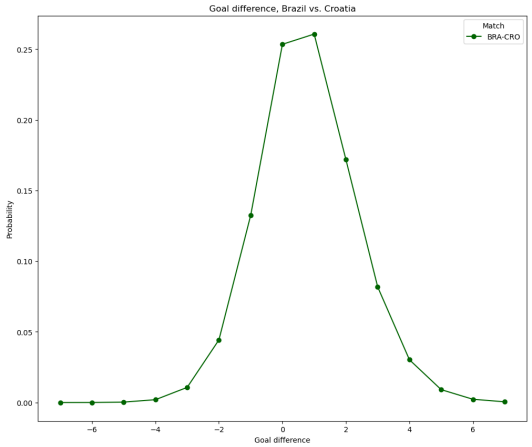

Figure 7: Predicted goal difference in Brazil vs. Croatia match

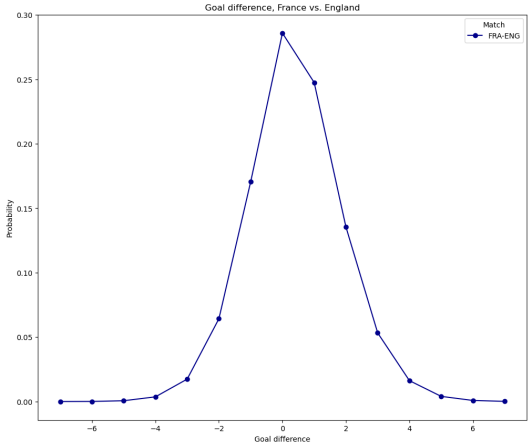

Figure 8: Predicted goal difference in France vs. England match

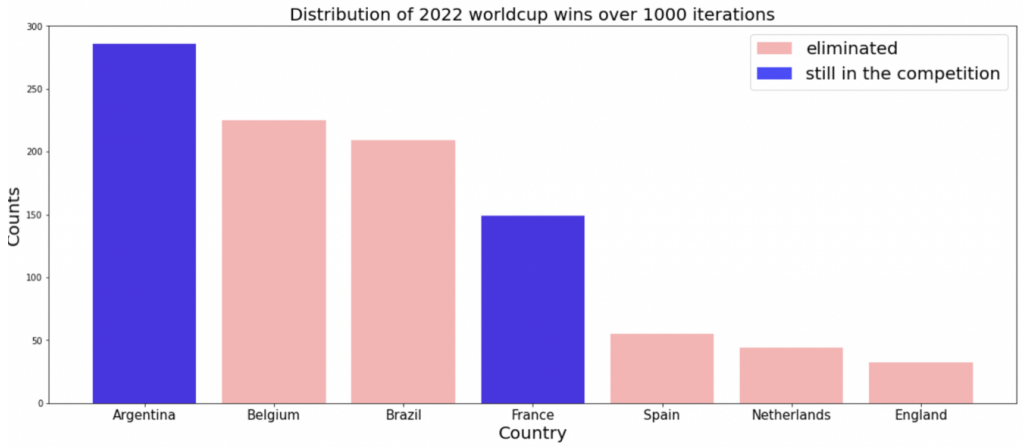

Figure 9: 1000 stochastic World Cup Predictions based on our logistic regression model fitted on 2018-2022 international games and FIFA rankings

- $\text{gc}_A^j \propto \sum_{i=1}^{j-1} \frac{1}{\gamma^{m_i}} \times$ (average number of goals conceded by team A in its i^{th} participation)

Our final score for team A in year m_t was simply $s_A^t = \text{gs}_A^t / \text{gs}_A^t$

Model	Train (%)	Validation (%)
Logistic	0.801	0.804
Single Decision Tree	0.827	0.766
Random Forest	0.823	0.793
Bagging	0.827	0.763
Adaboost	0.827	0.778

Table 3: Classification models' accuracy

6 Stochastic Simulations

In table 3, we could see that the logistic regression is the model performing the best on the validation set. We recall that the data set for these experiences stem from international games since the last worldcup in 2018, and from the FIFA ranking data set.

Our best model predicts Brazil to win the world cup. One must say that this is not a bad guess as Brazil won 5 world cups in the past, and is an objectively strong team.

However, we know that this is wrong as Brazil lost in the quarter-finals. We noticed as well that while this model was able to replicate our intuitive understanding of each team's relative strength, this model did not predict major upsets of the 2022 Qatar world cup. It did not foresee that Brazil would lose against Croatia, and that Spain and Portugal would lose against Morocco. Hence, we decided to include a stochastic element in our predictions that would give an advantage to the underdog of the game, at every game of the simulated tournament. After simulating 1000 worldcups, we obtained the following worldcup winner distributions 9. We can see that even though our model predicted Brazil to win the world cup, it favors in an ordered average Argentina, Belgium, Brazil France, Spain, Netherlands and England. Even though 5 of them are currently disqualified, Argentina and France are still in line to win the worldcup reflecting the quality of our model. However, it still gives no chance to Morocco.

7 Limitations and Future Work

We believe that one of the main problems to predict the World Cup is the reduced number of observation points for each pair of teams, and the fact that the tournament only happens every four years means that the likelihood that two teams will face each other again is not very high, particularly during the elimination stages. This makes our matrix sparse after performing one hot encoding over the home and away teams, and was the main reason that drove us to look into complementary datasets for country teams.

Another relevant issue is that it is difficult to characterize the qualitative player differences over the years, because we don't have time series data for the specific members (and their positions) that have participated in the previous tournaments. This could explain why the team level aggregation of the players' data set didn't help to improve the score of our models. One alternative here could be web scraping the exact players and positions for each team, as well as collect more information on the players themselves.

Finally, we would also like to keep exploring different models, such as the Double Poisson Regression.

8 References

Azhari, Hafis Rialdy, Yekti Widyaningsih, and Dian Lestari. "Predicting final result of football match using poisson regression model." Journal of Physics: Conference Series. Vol. 1108. No. 1. IOP Publishing, 2018.

Nguyen, Q. (2021). Poisson Modeling and Predicting English Premier League Goal Scoring. arXiv preprint arXiv:2105.09881.

Bruinsma, R. (2020). Using Poisson regression to model football scores and exploit inaccuracies in the online betting market (Doctoral dissertation).

Penn, Matthew J., and Christl A. Donnelly. "Analysis of a double Poisson model for predicting football results in Euro 2020." Plos one 17.5 (2022): e0268511.

Mart Jüriso, "Predicting FIFA 2022 World Cup with ML", https://www.kaggle.com/code/sslp23/predicting-fifa-2022-world-cup-with-ml