Emilio Zavala Miceli

3zavalam.tech@gmail.com

https://www.linkedin.com/in/emilio-zavala-miceli-86595927b/

January 2025

Predicting Expected Goals (xG) in Football: A Machine Learning Approach Using 2022 FIFA World Cup Data

1. Introduction

In this project, we developed a machine learning model to predict Expected Goals (xG) in football. The xG metric quantifies the likelihood of a goal being scored from a particular shot based on various factors such as shot location, angle, and body part. By leveraging historical data and advanced machine learning techniques, the model aims to provide more accurate predictions of goal-scoring opportunities, offering valuable insights for teams, analysts, and fans to better understand match dynamics and player performance. We used data of the 2022 World Cup (WC) provided by Hudle Statsbomb and retrieved thanks to the *mplsoccer* library.

2. Method

2.2 Data Collection

Statsbomb is a leading company in the sports data industry, particularly in football. In recent years, the company has released data from major tournaments. For this project, we focus on the 2022 FIFA World Cup (WC) in Qatar. While the size of the dataset is substantial, there is potential for more data from other competitions. However, it’s important to recognize that international football differs from club football, so any additional data should be from the same context to ensure consistency. The dataset of the WC 2022 includes nearly 4,000 events per match, where approximately 20 events are shots. In total, there are almost 1,500 shots in the tournament *(n=1,453)*.

We collect the data from statsbomb using the *mplsoccer* library. This allowed us to access all matches from the World Cup, retrieve their match id's, and extract the corresponding events for each match.

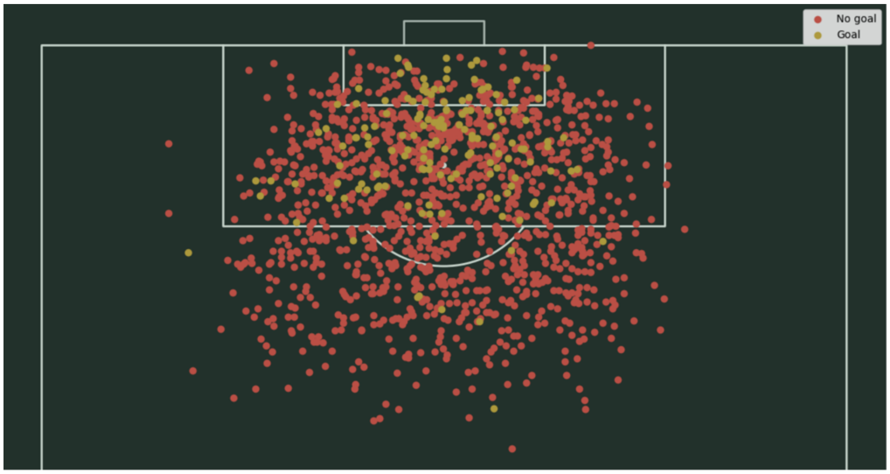


Figure 1 – All 2022 World Cup shots

To prepare the data, we created a new dataframe containing only the features required for training the model. The selected features, which we believe have the most significant impact, include:

- Location (x, y)
- Outcome (e.g., saved, goal, etc.)
- Type (e.g., open play, corner, etc.)
- Body part (e.g., right foot, left foot, head, etc.)
- Under pressure (Boolean)
- Technique (e.g., volley, normal, etc.)
- Statsbomb xG (a value provided by statsbomb's model)

By including the xG values from the more advanced statsbomb model, we can compare our results directly against their predictions.

Next, we filter the data to include only shots taken during the first four periods: the 1st, 2nd half, 1st extra time and 2nd extra time. This step is crucial because statsbomb data also includes shots from penalty shootouts, which could distort the model's understanding of typical match scenarios.

In football, the angel and distance from which a shot is taken are critical factors to determine the likelihood of scoring a goal. A wider shooting angle increases the chances of scoring, as the goalkeeper has more area to cover. Similarly, shots taken closer to the goal are generally more likely to result in goals due to the goalkeeper's reduced reaction time. So, we calculate manually both the angle and distance of each shot.

We transform categorical data into numerical form to prepare the dataset for model training. The technique named and type name columns are converted into dummy variables, creating separate columns for each unique value in those categories, with binary values indicating the presence of each technique or sub-type. The under-pressure column is first filled with 0 for missing values and then converted to integers (0 or 1) to represent whether a shot was taken under.

2.3 Data Analysis

To analyze the data, we used Altair to visualize the relationship between shot angle, distance, and the likelihood of scoring a goal. First, we plotted a scatter plot of shot angle versus goal outcome, then added a regression line to model the relationship. Similarly, we created a scatter plot for shot distance versus goal outcome and calculated the correlation between these features and the goal outcome. The correlation coefficients for both the shot angle and distance were computed to assess their impact on scoring. The correlation for distance was -0.25 and 0.31 for the angle, suggesting that the angle of a shot has a stronger influence on the chance of scoring than the distance from the goal.

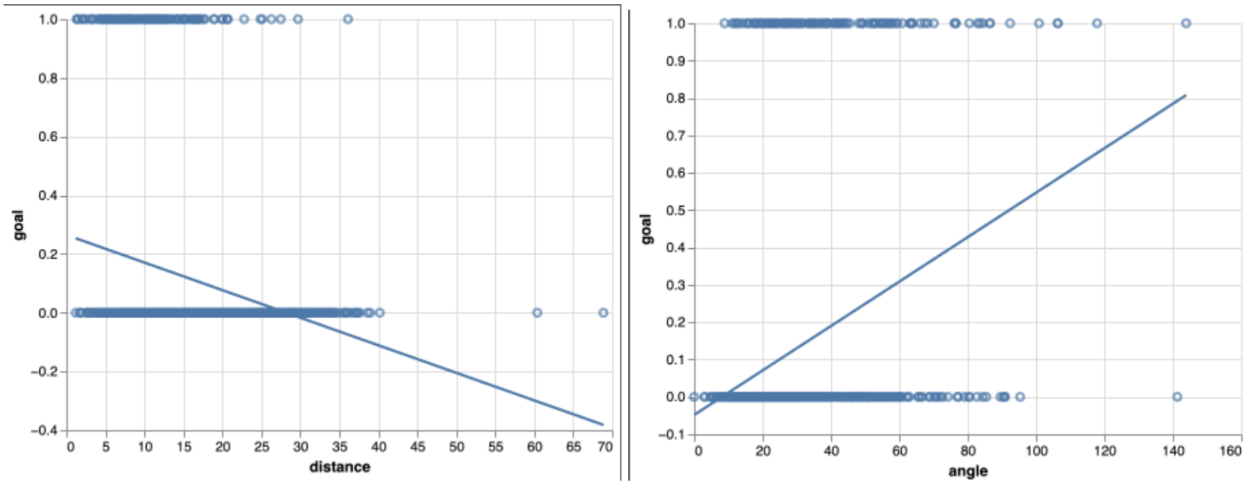


Figure 2 - Shot and Angle correlation with Goals

The first chart shows the relationship between shot distance and the probability of ending in a goal. The regression line slopes downward, indicatin gthat the probability of scoring decreases as the distance increases. Most shots, especially from long distaces, result in no goal, thoght a few rare outliers shows goals from farther away. This suggets that close-range shots are more effective. The second chart illustrates the relationship between shot angle and goal probability. The positive slope of the regression line shows that the likelihood of scoring increases with the angle. However, most data points are clusted at 0 (no goal), suggesting that other factors also play a key rol in scoring. The few goals from difficult angles highlight that while it’s possible to score from tough positions, it’s still uncommon.

2.4 Modelling

The first step is to select the model we want to use. In this case, we will train both Linear and Logistic regression using the *sklearn* library to determine which one performs best. For simplicity, we will use only distance an angle as our features (X). We will trian both models and evaluate their performance by calculating the R^2 score to see how each model responds to the data.

The R^2 score of the linear model is 0.099, and for the logistic model is 0.096. These low values suggest that neither model expalin much of the variance in the data, meaming the predictions don't closely match the actual outcome. In comparison, the statsbomb xG model has a score of 0.20.

To chose the best model, we need to dig deeper into the analysis. First, the linear regression model produces at least 45 neagative xG values, which is not possible because expected goals cannot be negative regardless of shot's distance or angle. In figure 3, in the left one we can see the relationship between shot angle and the xG predicted by the linear model. The right one shoes the anlge again, helping to understand the modls prediction pattern and how the shot angle correlates with the expected goal value. The negative values in the linear regression model are a red flag, indicating that his model is not appropiate for predicting xG, as xG values should always fall between 0 and 1. Since the model produces impoosibel values, we cannot rely on it for this type of prediction, and thus it is not suitable for our analysis.

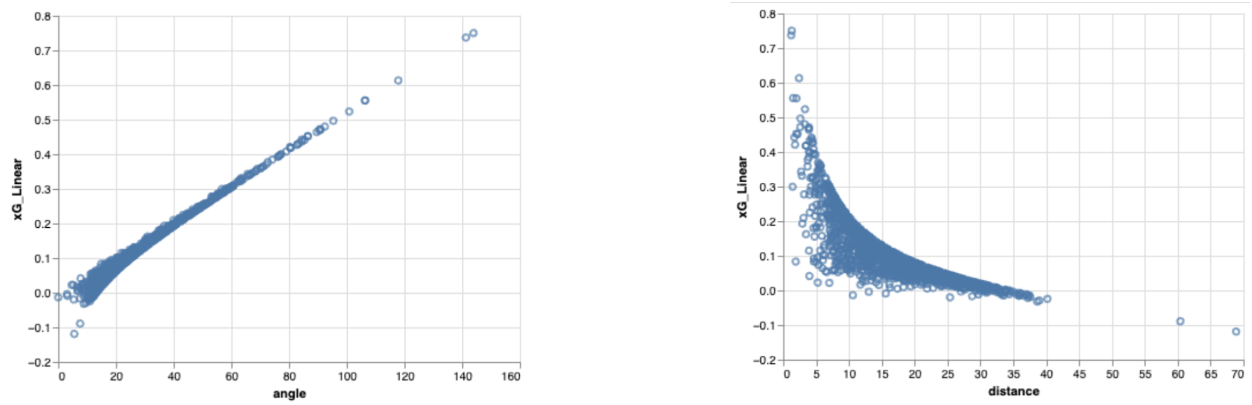


Figure 3 - Angle and Distance relation with Linear Regression

For the logistic regression model, we don't encounter negative values, which is a postive aspect. Next, we analyze the relationship between angle and distance to asses if ther eis a correlation, which will help us undeerstand how these features contribute to the model’s prediction of xG.

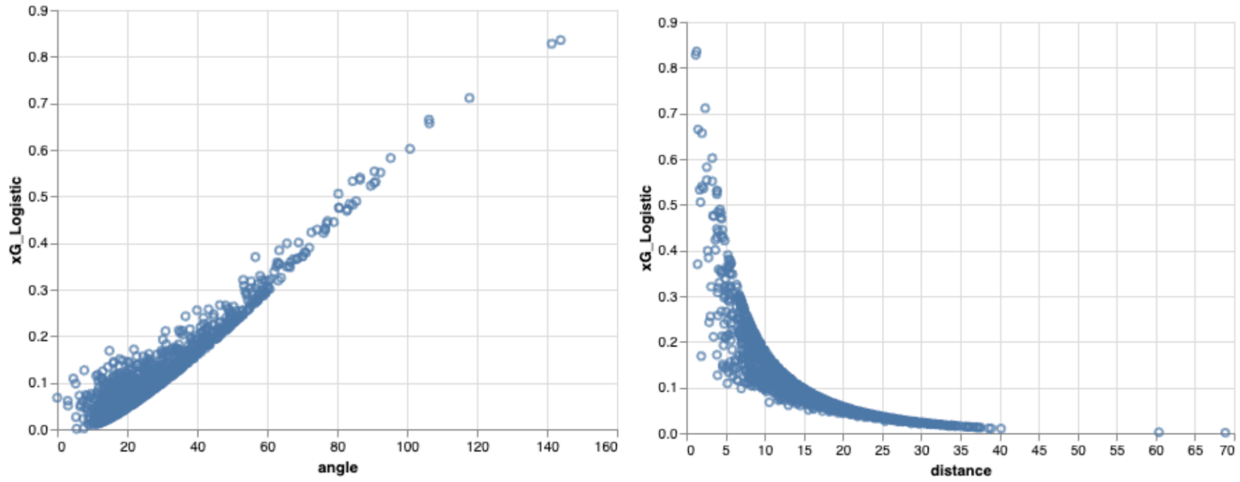


Figure 4 - Angle and Distance relation with Logistic Regression

Now that we choose the model we are using, is time to selected a set of features from the dataaset to use as input variables for the model. Th features chosen include factors such as wheter the shot wwas taken under pressure, shot angle, distance from the goal, the player’s preferred side, whether the shot was a header, and different types of techniques and shot scenarios like open play or penalties. We created a new subset of the data containing only these selected features, which will be used to train the model. The first few rows of this subset were displayed to ensure the data was properly preapred for the next steps in the analysis.

In this analysis, the target variable we are trying to predict is whether the shot resulted in a goal or not. This variable, labeled as ‘goal’, takes a value of 1 if the shot was a goal and 0 if it wasn’t. It represents the outocme of the shot, which we aim to predict based on the features we selected earlier.

The code trains a logistic regression model using the selected features (X) to predict the likelihood of a shot resulting in a goal (y). By applying the predict_proba(X) method, the model generates predicted probabilities for both classes (goal and no goal). Selecting [:, 1] extracts the probability of the shot being a goal. With the addition of these new features, the model's performance improves, as evidenced by an increase in the R^2 score to 0.19. This indicates a notable enhancement in the model's ability to explain the variance in the dataset.

3. Results

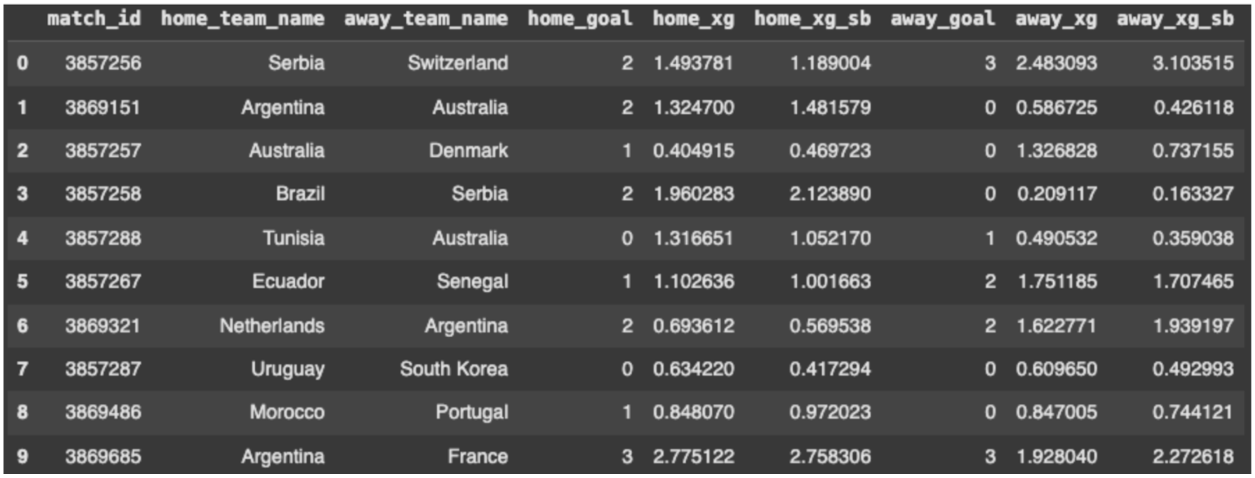
3.1 Evaluating in All Matches

To evaluate all matches and calculate the xG for each shot, we followed a structured approach. The first step involved preparint the data. We created a dataframe, which contains essential match details suach as the match id, the home and away names. Next, we looped through each amtch using a loop over the dataframe. For every match, we retrieved the event data that includes all shot attempts. This was achieved by filtering for shot events for each match id. This gave us a dataset focused solely on shots taken during the match.

To ensure we considered only relevant data, we applied an additional filter to focus exclusively on shots taken during open play. This step ensured that our analysis was based on regular gameplay, aligning with standard xG calculations.

With the filtered data, we proceeded to calculate the xG for each shot using a custom function. This function incorporates various shot features, such as angle, distance, body part used, technique, and shot type. These features are passed into the logistic regression model, which predicts the likelihood of the shot resulting in a goal, providing an xG value for each attempt.

After calculating the xG for each shot, we separeted the shot by the home and away teams. For each tema, we calculated three key statistics: number of goals scored, total xG, and the statsbombs xG. This approach allows us to assess the performance of the xG model for each shot across all matches, providing insights into the model's accuracy and highlighting any discrepancies between expected and acutal goals scored.



	match_id	home_team_name	away_team_name	home_goal	home_xg	home_xg_sb	away_goal	away_xg	away_xg_sb
0	3857256	Serbia	Switzerland	2	1.493781	1.189004	3	2.483093	3.103515
1	3869151	Argentina	Australia	2	1.324700	1.481579	0	0.586725	0.426118
2	3857257	Australia	Denmark	1	0.404915	0.469723	0	1.326828	0.737155
3	3857258	Brazil	Serbia	2	1.960283	2.123890	0	0.209117	0.163327
4	3857288	Tunisia	Australia	0	1.316651	1.052170	1	0.490532	0.359038
5	3857267	Ecuador	Senegal	1	1.102636	1.001663	2	1.751185	1.707465
6	3869321	Netherlands	Argentina	2	0.693612	0.569538	2	1.622771	1.939197
7	3857287	Uruguay	South Korea	0	0.634220	0.417294	0	0.609650	0.492993
8	3869486	Morocco	Portugal	1	0.848070	0.972023	0	0.847005	0.744121
9	3869685	Argentina	France	3	2.775122	2.758306	3	1.928040	2.272618

Table 1 - Matches Comparison

3.2 Results Analysis

The results of the analysis demonstrate the performance of the custo mlogistic regression model for predicting expected goals and its comparison with statsbomb xG values. The model produced realistic xG predicitons, as it did not generate negative values, which would be illogical in the context of football. By analyzing the total goals scored and comparing the total xG values for both teams across all matches, it became clear that while the model showed promise, there were occasional discrepancies between the predicted xG and actual goals .

When comparing the model's predict xG to the statsbombs xG values, we observed that statsbomb predictions generally aligned more closely with the actual goals scored. This indicates that statsbomb model may have captured more complex nuances of shot outcome that our simplier logistic regression model might have missed, such as defensive pressure or player positioning. The custom model, while fairly accurate, showed minor over- or under-predictions. This could be attributed to the limitations of the features used during training or the relatively small dataset.

In terms of model accuracy, using metrics like total xG per team and the number of goals scored, our model performed reasonably well but showed a slight gap when compared to statsbomb more refined approach. These differences suggest areas for improvement, such as including additional features (e.g., shot velocity, match context, player positioning) and tuning the model further.

4. Conclusion

Overall, the custom logistic regression model provided a solid starting point for predicting expected goals (xG) based on shot features. However, there is significant room for enhancement. The comparison with StatsBomb's xG model highlights the strengths of the custom model while also identifying areas for improvement. In particular, incorporating more granular features related to the shot (e.g., defensive pressure, player positioning, shot technique) could enhance the model's predictive power.

While the custom model shows potential, StatsBomb's xG, which benefits from a much larger dataset and more sophisticated algorithms, performed better in terms of predicting actual goal outcomes. The custom model's ability to avoid negative xG values is an advantage, but further development is necessary to match the accuracy and reliability of StatsBomb.

In conclusion, while the model is a promising tool for evaluating shot quality and providing insights into team and player performance, it still has limitations that need to be addressed. Future improvements to the model, including more complex feature sets and advanced machine learning techniques, would enhance its ability to predict goals and provide more accurate insights for football analysis.

Bibliography

Durgapal, A., & Rowlinson, A. (2021). *mplsoccer* (Version 1.4.0) [Software]. https://github.com/andrewRowlinson/mplsoccer