xSFR - Expected Shot From Recovery

Technical report

Jedrzej Sarna

October 2024

1 Introduction

This report introduces a new metric **xSFR**, which defines the likelihood of a shot occurring within 10 seconds after regaining possession. It focuses on the technical aspects of the metric, providing a detailed explanation of the process involved in developing the model behind it.

The purpose of the metric is to understand how a team transitions from ball recovery to shot creation, as well as to evaluate individual actions that can lead to goal-scoring opportunities. Additionally, the metric enables an assessment of each player's impact during the crucial initial moments after regaining possession, identifying who performs best in this role.

2 Model Theory

The metric was formulated as a machine learning problem. More specifically, based on data from the 2017/18 Premier League season, a model was developed to identify the factors influencing the likelihood of a shot occurring after regaining possession. A logistic regression model was chosen, where the output can be interpreted as the probability of the desired event. Logistic regression is particularly useful because it allows for easy interpretation of the impact of individual explanatory variables on the dependent variable. Therefore, the problem can be defined mathematically in following way: the probability of the event $y = 1$ (a shot occurring within 10 seconds after recovery) is modeled as a function of the explanatory variables X. The relationship between the explanatory variables and the probability is expressed using the logistic function:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

where:

- $P(y = 1|X)$ is the probability of a shot occurring within 10 seconds after recovery,
- β_0 is the intercept,
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the explanatory variables $X_1, X_2, \ldots, X_n$,
- e is the base of the natural logarithm.

The goal is to estimate the coefficients $\beta_0, \beta_1, \ldots, \beta_n$, which will be done using maximum likelihood estimation (MLE), hence maximizing the function

$$L(\beta_0, \beta_1, \ldots, \beta_n) = \prod_{i=1}^{N} P(y_i|X_i)^{y_i} (1 - P(y_i|X_i))^{1-y_i},$$

with respect to the coefficients. This yields the optimal parameter values that best fit the data.

3 Data Preprocessing & Analysis

The problem falls within the realm of machine learning, requiring a clearly defined output and input. Therefore, it is essential to delve deeper into the data preparation process, which is key for ensuring that the data used by the model is appropriately structured and relevant. The data used to build the model is event data sourced from Wyscout, covering the 2017/18 Premier League season.

3.1 Dependent Variable

To define the dependent variable based on the available data, we must first precisely define the concept of ”recovery.” In the model, it is assumed that a team has recovered the ball when, in the preceding events, the opposing team was in controlled possession (was able to make a pass or take a shot), and subsequently, the team in question gains control and performs a pass or shot. This definition allows us to isolate the specific aspect of play we are interested in, excluding situations where, for instance, there is a contest for a loose ball that lasts an extended period.

Next, it is important to consider the selection of the time frame within which a shot is required. Although the 10-second window is not explicitly mentioned in the name of the metric, it is a crucial aspect. This choice is justified by the analysis of available data and domain knowledge. Analyzing the previous Premier League season shows that only 2% of recoveries result in a shot within 5 seconds, 5% within 10 seconds, and over 7% within 15 seconds.

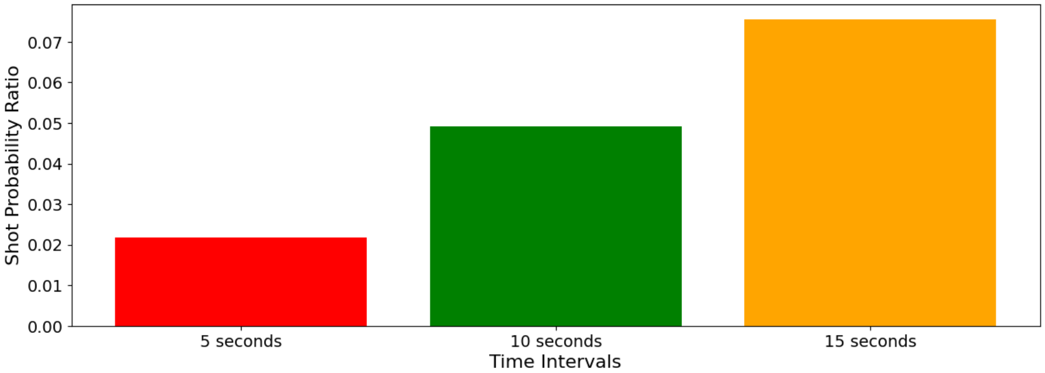


Figure 1: Shot probability after recovery within different time interval in Premier League 2017/18 season

Additionally, based on *Soccermatics* definiton of possession chain, average duration of one for a team is around 9.3 seconds, reinforcing the idea that a 10-second interval is appropriate. Considering longer periods may not be appropriate, as the event of a shot will not be closely related to the initial moment of ball recovery, which is the key event analyzed within the metric. Using this approach, we identify 25,972 relevant events (recoveries), of which 1,277 resulted in a shot within 10 seconds.

3.2 Independent Variables

Moving on to the independent variables, many factors were considered that could influence the probability of a shot following a recovery. The report includes the final selected variables and the methods used to prepare them. The selection of variables was primarily based on domain knowledge and the format of the available data.

Distance to goal
The variable represents the distance to the opponent’s goal, calculated based on the coordinates of the ball recovery location.

Minute of match
The variable represents the minute of the match. In the original Wyscout data, the timing of events on the field is recorded in seconds and halves. For the purposes of the model, this was converted into minutes to improve the interpretability of the final variable.

Pass distance
The variable represents the length of the first pass made by the player after recovering the ball.

Anticipation or interception
Indicates whether the player regained possession through anticipation or interception, as opposed to other methods such as a sliding tackle or less standard recoveries. These two attributes were combined due to the low frequency of anticipation events in the dataset and their similar characteristics, based on the Wyscout definitions. A value of 0 indicates another type of recovery (such as a sliding tackle or a less standard action). Anticipation and interception were also tested separately in the model, however they did not yield positive results.

Result
The current result of the match. The logic can be compared to a "balance sheet." If a team scores a goal, the Result variable for that team is assigned a value of +1, while the opposing team receives a value of -1. In the case of an equalizing goal, the Result variable for both teams is set to 0, and the logic continues.

Central area
Indicates whether the ball recovery occurred in a central area of the pitch. The variable defines the player's position relative to the width of the field. The central area is designated as the width of the penalty box, so a value of 0 represents the wide areas, also known as "channels." Despite the presence of a similar variable, Distance to goal, this variable provides additional information regarding the spatial dynamics of the recovery.

Below is a summary table of the value ranges for each variable, along with the correlation matrix.

	Distance to Goal	**Minute of Match**	**Pass Distance**	**Anticipation Interception**	**Result**	**Central Area**
mean	71.32	46.28	22.91	0.20	0.02	0.66
std	18.45	26.96	16.53	0.40	1.16	0.47
min	3.00	0.05	0.00	0.00	-6.00	0.00
25%	58.55	23.29	11.40	0.00	0.00	0.00
50%	73.16	46.03	19.31	0.00	0.00	1.00
75%	86.27	69.01	30.41	0.00	1.00	1.00
max	110.46	102.11	133.66	1.00	6.00	1.00

Table 1: Summary statistics for variables used in the final model

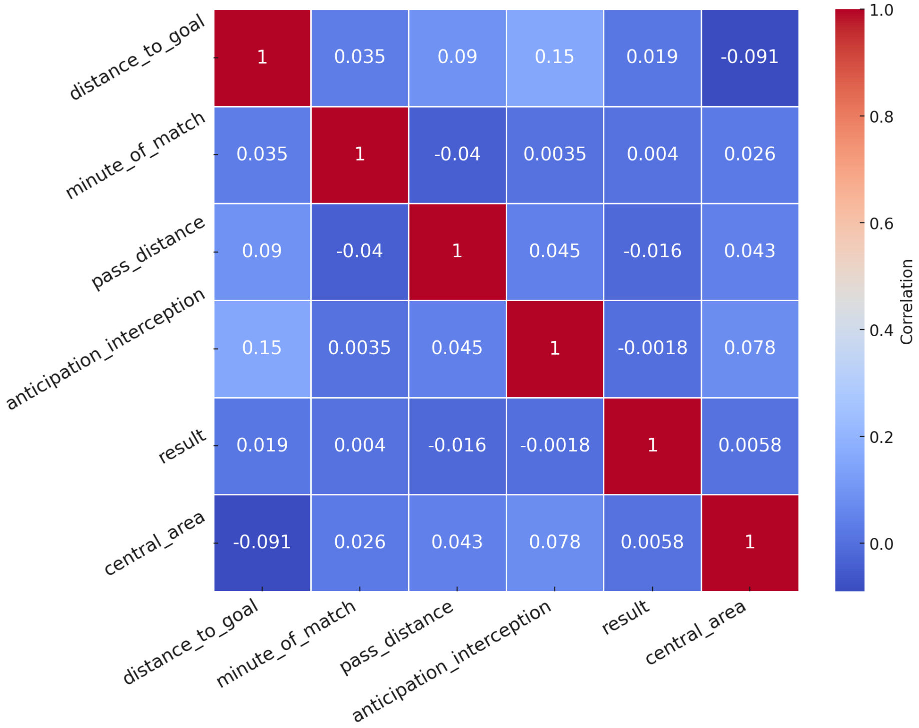


Figure 2: Correlation matrix for variables used in final model

As can be observed, the presented variables are not highly correlated with each other, which is desirable in a logistic regression model. Each variable describes a different aspect of the game, ensuring that they contribute unique information to the model. Lastly, Variance Inflation Factor (VIF) was considered, which helps determine the level of multicollinearity between the independent variables in the model. As can be seen in Table 2, all predictors (excluding the constant) have VIF values close to 1, indicating no issues with multicollinearity in the model.

Feature	VIF
const	22.517
distance_to_goal	1.046
minute_of_match	1.004
pass_distance	1.014
anticipation_interception	1.034
result	1.001
central_area	1.021

Table 2: Variance Inflation Factor (VIF) for Model Features

4 Building Model

After defining and analyzing the explanatory variables, we can proceed to building the logistic regression model. Training the model on event data from the 2017/18 Premier League season results in a model defined by the following equation

$$\begin{aligned}\operatorname{logit}(P(y=1|X)) = &-0.569 - 0.041 \cdot \mathit{distance\_to\_goal} + 0.004 \cdot \mathit{minute\_of\_match} \\ &- 0.005 \cdot \mathit{pass\_distance} + 0.369 \cdot \mathit{anticipation\_interception} \\ &+ 0.082 \cdot \mathit{result} + 0.153 \cdot \mathit{central\_area}.\end{aligned}$$

The complete summary of the model can be obtained through a results table, which provides detailed insights into the performance and coefficients of the model.

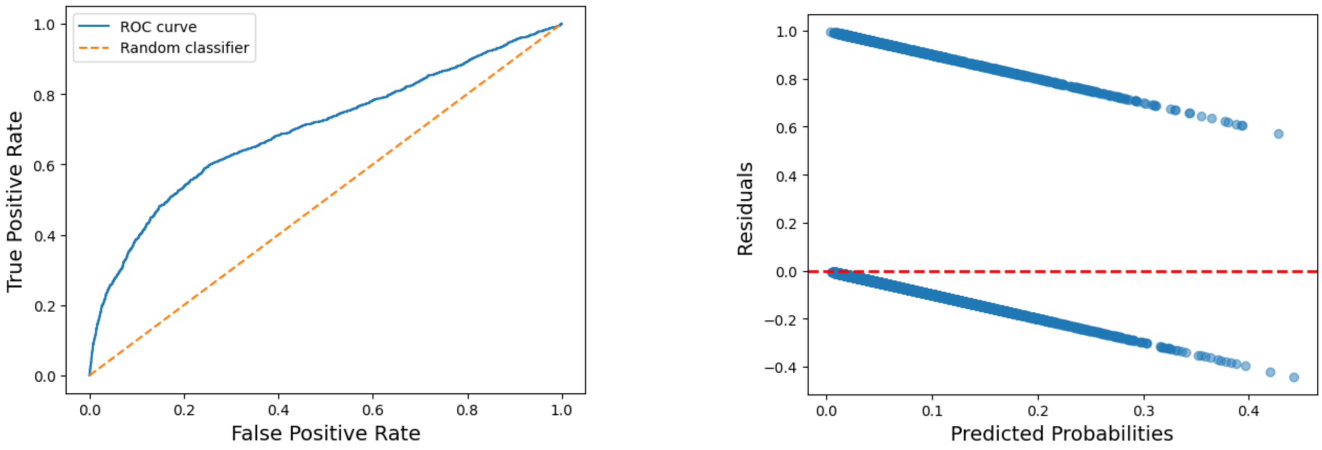
Variable	Coef.	Std. Err.	z	P > |z|	[0.025, 0.975]
const	-0.569	0.126	-4.517	0.000	[-0.815, -0.322]
distance_to_goal	-0.041	0.002	-26.759	0.000	[-0.044, -0.038]
minute_of_match	0.004	0.001	3.305	0.001	[0.001, 0.006]
pass_distance	-0.005	0.002	-2.763	0.006	[-0.009, -0.002]
anticipation_interception	0.369	0.072	5.153	0.000	[0.228, 0.509]
result	0.082	0.024	3.361	0.001	[0.034, 0.129]
central_area	0.153	0.067	2.303	0.021	[0.023, 0.284]

Table 3: Logistic Regression Results

It is important to note that the described model is the final version. During the development process, many other explanatory variables were considered (including whether the first pass after recovery was forward or backward, the speed of the first pass, etc.), but they were ultimately excluded as they were statistically insignificant in the model. Therefore, as can be seen in the results table, all the remaining variables are statistically significant, with p-values less than the commonly accepted threshold of 0.05.

5 Evaluating Model

Let us now focus on the quality of the model itself. Considering the R^2 value, the model achieves a moderate score of 70%. Additionally, the AUC-ROC score of 0.699 suggests a reasonably good ability to distinguish between positive and negative outcomes.



(a)

(b)

Figure 3: The ROC curve (a) and the residuals plot (b).

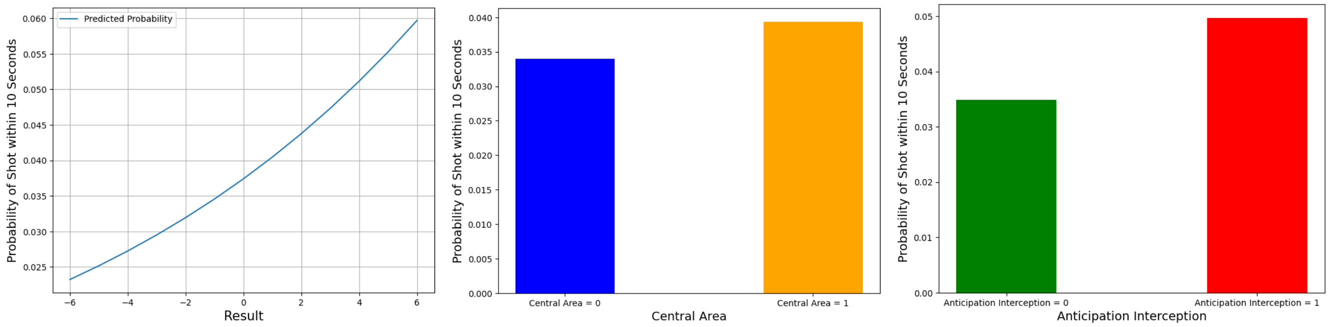
Analyzing the residuals from Figure 3, it can be seen that they are clustered mostly on the left side (lower predicted probabilities). This indicates that the model is more confident about predicting no shots than predicting shots, which is also reflected in Table 4. This is also unfortunately caused by the class imbalance in predicted variable.

Class	**Correct (True)**	**Incorrect (False)**
Shot Class	0.625	0.375
No Shot Class	0.700	0.300

Table 4: Proportion of Correct and Incorrect Predictions for Shot and No Shot Classes

Although the model does not achieve exemplary performance, it demonstrates good predictive accuracy. During the analysis, additional variables were considered to improve the model, but it was not possible to enhance it without overfitting. At this point, it is worth considering whether the challenge lies in the nature of the problem itself. It may be that predicting the occurrence of a shot based on a ball recovery and related events is not as straightforward as initially thought. For this type of problem, access to tracking data could also be beneficial, as it would provide insight into the positioning of players, adding valuable information. However, this falls outside the scope of this project.

Despite these limitations, the model does have predictive capabilities, and it is worth delving deeper into the interpretation of the impact of individual variables. By selecting one variable and replacing the rest with their average values, we can analyze how individual variables impact the dependent variable. While the estimated coefficient for each variable already provides some insight, this additional analysis offers a deeper understanding of each variable's effect.



(a)

(b)

(c)

Figure 4: Change in dependent variable based on Result (a), Central Area (b) and Anticipation or Interception (c).

Above, the plots for the three variables with the greatest impact on the dependent variable are presented. It is also important to note, that within the model, the estimated coefficients align with football intuition. Going through them one by one, we observe:

- **Distance to goal (coefficient: -0.041)**: Shorter distance increases the likelihood of a shot, as teams are in a more advantageous position to attempt a quick attack.
- **Minute of match (coefficient: 0.004)**: Later stages of the game see slight increase in the frequency of shots following recoveries, which might be due to teams taking more risks.
- **Pass distance (coefficient: -0.005)**: Longer passes may result from a sense of urgency and can often be inaccurate, whereas a short pass to a nearby teammate can prove to be a more effective option.
- **Anticipation or interception (coefficient: 0.369)**: Anticipation or interception has a significant impact as the player has better control of the ball, being able to make more deliberate decisions about next actions.
- **Result (coefficient: 0.082)**: While not directly affecting the immediate chance of a shot, this could also describe team strength and dominance on the pitch.
- **Central area (coefficient: 0.153)**: Recoveries in central areas are more likely to lead to a shot due to the strategic advantage of these positions in launching attacks.

6 Recommended Players by Model

After completing the training process, the next step involved testing the model on the remaining top 5 leagues and identifying the best-performing players. To ensure a fair comparison, it was necessary to account for the number of minutes played, so normalization per 90 minutes was applied. Additionally, only players who played more than 60% of the total minutes possible in the previous season were considered, which equates to over 2052 minutes. The top-performing players, according to this metric, achieved values ranging from 25% to even 40%, representing a significant improvement over the league average. Unsurprisingly, defensive midfielders, who are primarily responsible for ball recoveries, dominated the rankings. The table below presents the top 5 players in each league according to the metric.

League	**Ligue 1**		**La Liga**		**Serie A**		**Bundesliga**	
Ranking	**Name**	**xSFR**	**Name**	**xSFR**	**Name**	**xSFR**	**Name**	**xSFR**
1	F.Anguissa	0.40	G.Kondogbia	0.32	Allan	0.32	M.Hummels	0.34
2	A.Toure	0.29	J.Lerma	0.30	J.Palomino	0.29	J.P.Gbamin	0.29
3	T.Ndombele	0.29	Casemiro	0.30	L.Leiva	0.29	P.Skjelbred	0.29
4	Fabinho	0.27	F.Ruiz	0.29	T.Rincon	0.28	D.Latza	0.27
5	B.Andre	0.27	M.San Jose	0.29	M.Badelj	0.28	C.Aranguiz	0.26

Table 5: Top 5 players in each league based on xSFR performance.

The player who stands out significantly in terms of xSFR is Frank Anguissa, who had an outstanding season with Olympique Marseille. Based on this metric, he is the top recommendation.

7 Summary

In conclusion, although the model does not achieve outstanding results in terms of quality, it still demonstrates satisfactory predictive ability. The selected variables intuitively influence the likelihood of a shot occurring, and the standout players identified by this metric also appear to fit the expected profile. While the model could be further refined by incorporating additional variables, particularly those related to tracking data, overall, it serves as a useful tool for identifying potential future transfers for the club.