Data Science Exam
How events in football matches influence expected goals

MSc Cognitive Science Spring 2022

School of Communication and Culture

Lecturer: Chris Mathys

Rokas Maksevičius

Student number: 202103222

Character count: 25 653

Introduction

Data analytical approaches in sports have received wide attention over the last two decades. The well-known story of Billy Beane and the Oakland Athletics of 2002, which had a book published in 2004 (Lewis, 2004) and a film released in 2011, shows a team taking a statistical approach in selecting the players and strategies used, outperforming teams with significantly bigger budgets in this way. The central tenet is that there are two statistically significant attributes for a good player: the on-base percentage (how frequently a player reaches base) and the slugging percentage (total bases divided by at bats). Other statistical measures, traditionally thought to be significant, such as batting average or stolen bases, did not in fact have a lot of influence on the outcome of games.

Besides baseball, other sports analysts began looking at games more as a mathematical problem than one based on "intuition" or "instinct". For example, statistical analysis on basketball games (Zuccolotto et al., 2018) has been used to identify which factors create "high-pressure game situations" which affect how players shoot the ball. In ice hockey, as well as other sports with drafting systems, drafting strategies have been greatly improved using statistics of rookies who might not be familiar to the scouts otherwise (Nandakumar & Jensen, 2019).

In association football, a big breakthrough was the expected goals (commonly referred to as xG) model initially created and published by Sam Green (2012), a data scientist at Opta Sports. In the system, factors such as where the shot was taken from, where (in the goal) it was directed towards, what body part was used to shoot, and other metrics were used to quantify how good a certain chance is. This quantification was named Expected Goals, and, ranging from 0 to 1, tells how likely a player *should* be scoring given the circumstances of the event. Over time, other similar statistics were developed, ranging from expected assists (xA), a statistic with a similar premise as xG, to Post-Shot Expected Goals minus Goals Allowed (PSxG+/-) (Goodman, 2018), a statistic measuring how good a goalkeeper performs.

Besides the obvious intuition of telling players what type of chances they should try to create more, xG became a good tool of evaluating both teams and individual players. Due to the low-scoring nature of football, it is not necessarily obvious which team performed better when looking at a score line such as 0-0. Using expected goals, one can see how many viable chances a team generated, even if they were not successfully used. On the other side, xG can be used on individual players to estimate how good of a "finisher" they are. It is usual to find that world class players generally score 1.5 to 2 times more than what they were "supposed to" according to xG, because the models are fitted on data from multiple players, teams, leagues and seasons.

Because expected goals models are generally created within the industry by either sports journalism companies or the analytical department of a professional team, they are not openly available to use. Due to this, there is no exact definition of which parameters should be used for evaluation. For example, Opta Analytics, the company responsible for the initial xG model, boasts about using "up to 35" different contextual factors in their newest model. This potentially includes not only factors for the exact moment of the shot, but also factors leading up to the shot, such as the style of the pass, the distance the player had to run before taking a shot, the time of the event, and so on.

However, there are also completely external factors to consider: if a team has a player sent off, does that influence xG for the rest of the game? Does a missed penalty (quantitatively) demoralize a team? How do early substitutions impact the game? This paper will attempt to answer these questions by building a modified expected goals model and applying it to matches to see how the quality of chances changes when certain events within a game happen.

Methods

The analysis was carried out using Python 3.10, with matplotlib (Hunter, 2007) used to generate figures, scikit-learn (Pedregosa et al., 2011) for the regression models and their evaluation, and SciPy (Virtanen et al., 2020) for statistical analysis. The code for this project is available on GitHub: https://github.com/rmxrmx/ds_exam .

Dataset

The dataset used in this paper is "Football Events", available on Kaggle (Secareanu, 2017). It contains over 940 000 events from over 9 000 football games, taken from the strongest 5 leagues (England, Germany, Spain, Italy, France) over 6 seasons from 2011 to 2017. There is a range of events in the set: shot attempts, corners, fouls, yellow and red cards, offsides and so on. Each event also has an "event team" and an "event player", and, optionally, a second player (for events such as substitutions, or goals where an assist happened).

Additionally, shots have more specific parameters: they have a *shot place* (such as "top left corner", "centre of the goal" or "misses to the right"), a *shot outcome* ("on target", "off target", "blocked", "hit the bar"), a wide range of *locations* ("right side of the box", "penalty spot", "long range", and so on), a *body part* ("right foot", "left foot", or "head"), an *assist method* ("none", "pass", "cross", and similar), the *situation* ("open play", "set piece", "corner", or "free kick") and a boolean parameter *fast break*. Finally, it has a boolean parameter *is goal*, which simply says whether this shot was a goal.

The events have both a time (the minute in the game, ranging 0-100) and a *sort order*, which says how the events are ordered within a single match. For example, the 5th event in a match would have a sort order of 5. There are a few things to note with regards to this. As mentioned before, the time parameter can range from 0 to 100; this is unusual, because a football game lasts for 90 minutes, as well as some additional time (set by the referee according to how much downtime there was over the course of the game). It could be that there were different methods of marking time; where one match might have an event on the 95th minute recorded as 90+5 (as is usually seen when watching a match on television), another might have left it as 95. As can be seen in Figure 1, while the number of events after 90 minutes is non-zero, it falls off a lot.

Moreover, there are two big outliers in the dataset, occuring at minutes 45 and 90. These are likely to be due to the additional time mentioned before; it is probable that every event that happened in the added time was simply attributed to the 45th (if the event happened in the first half) or the 90th (if the event happened in the second half) minute. The fact that there are about twice as many events attributed to minute 90 as compared to minute 45 can be explained by the fact that there is generally more time added (and thus more time for events attributed to minute 90) at the end of the second half compared to the first; this, in turn, is due to more downtime in

the second half (see Figure 2; according to typical rules of the game, each substitution should increase the added time by about 30 seconds). Due to this reasoning, the outliers are left as-is for most of the analysis.

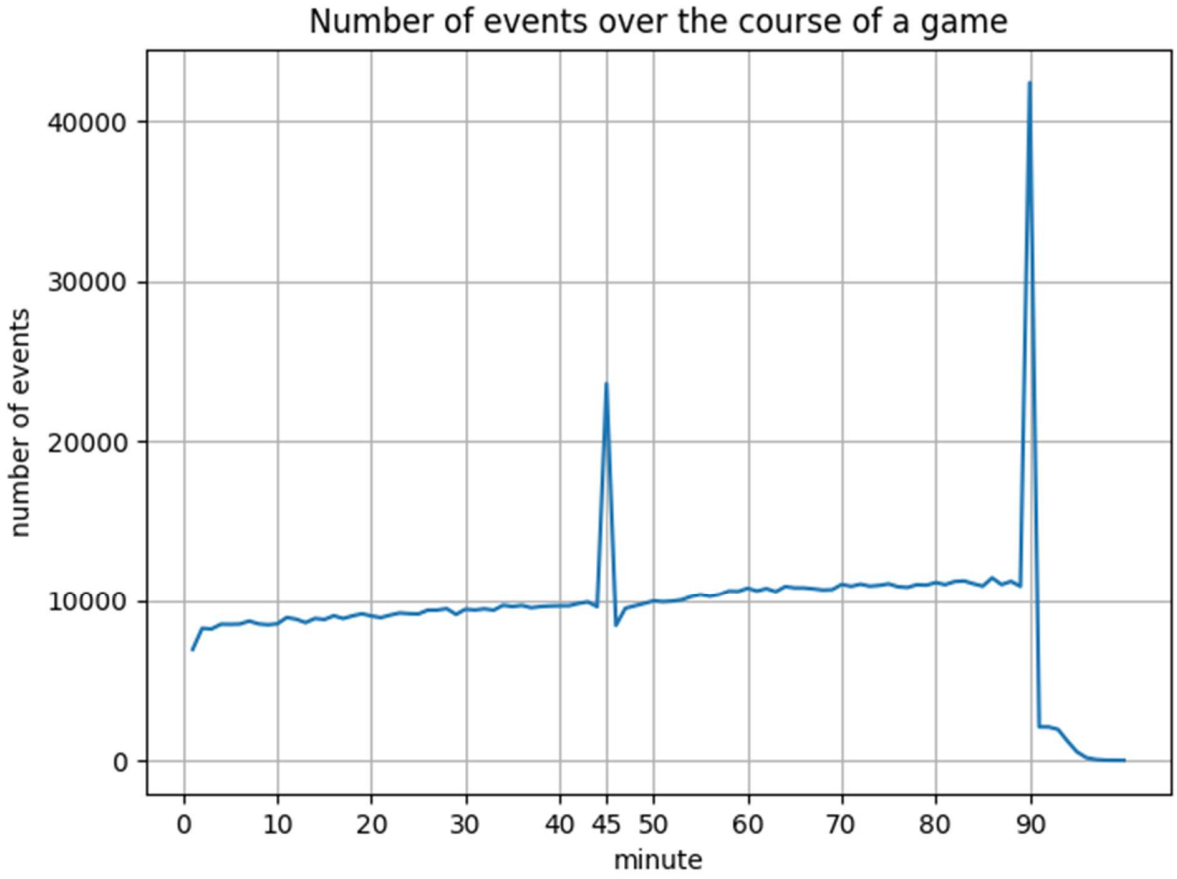

Figure 1: Number of events per minute in the dataset.

Expected goals model

The expected goals (xG) model was implemented by fitting a simple logistic regression model. The features for predicting whether a shot was a goal were *shot place, location*, *body part*, *assist method*, *situation* and *fast break*. The LR model was trained on 75% of the shot data and tested on the other 25%.

An important factor is understanding the build and usage of the model. The model was not built to evaluate shots in real time; it was made to evaluate how good a certain chance was within the context of a past game. Due to this, *shot place* was included as a feature; because this parameter can take on values such as “too high”, “blocked”, or “misses to the left”, it would not make sense to use it in real-time (as one cannot know whether a shot will miss before it happens). Due to the inclusion of this parameter, the model will likely have a misinformatively high accuracy, as any shot where the *shot place* is some form of *miss* will be easily classified as a non-goal. However,

because this project does not attempt to predict whether a shot will be a goal, the false accuracy is acceptable.

On the other hand, the choice of a logistic regression model should influence the accuracy of the model negatively. This is due to the fact that while we are predicting how *good* a chance was, the model is categorical; this means that a chance of 0.49 will be classified as a non-goal every time, even though the average player should score a goal about every second time in an identical position. While shots with an xG similar to that should not be a big part of the dataset, it should still be kept in mind.

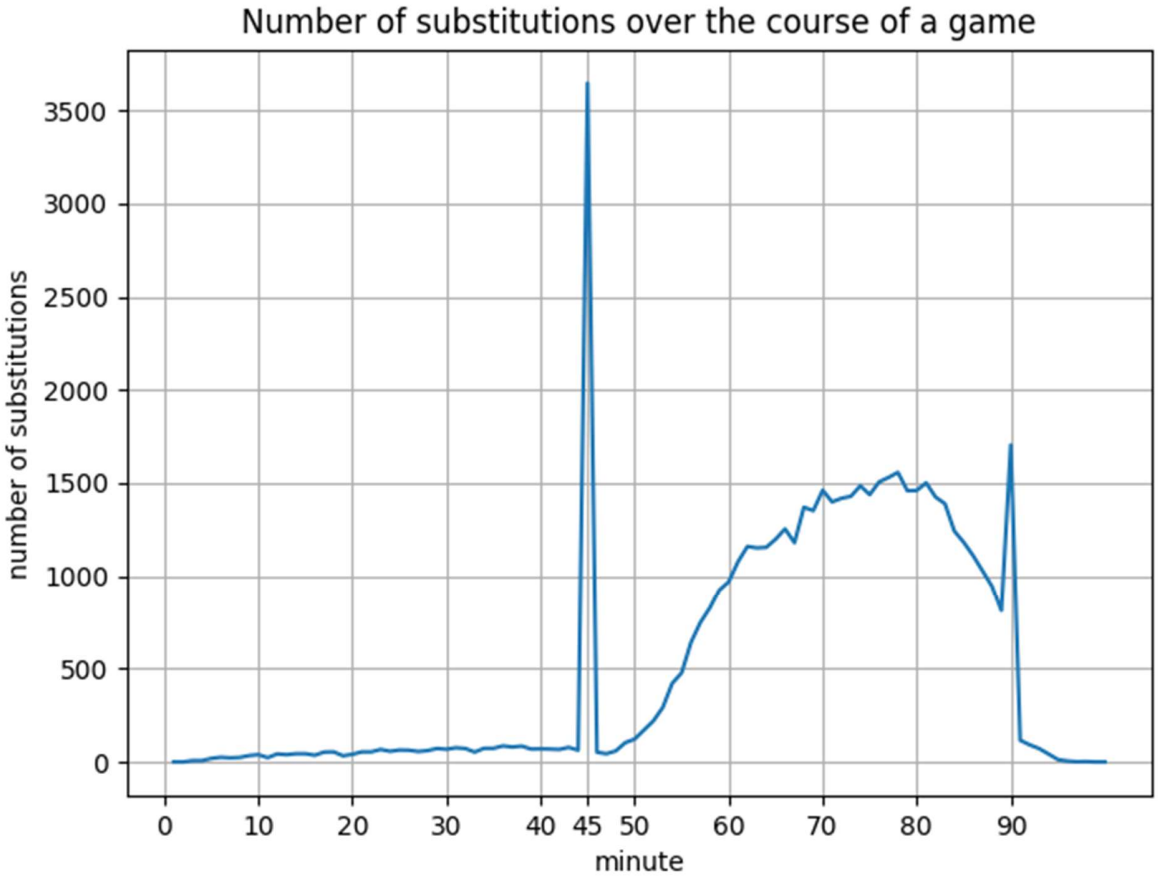

Figure 2: Number of substitutions per minute in the dataset.

High-impact events

The focal point of this project is to examine how a significant event within a game can change the rest of the game. To evaluate this, a few high-impact events were chosen: a sending off of a player (which means that one team plays with 1 player fewer for the rest of the game), a penalty kick (which has quite a bigger shot-to-goal ratio than an average shot - 76% of penalties are scored in this dataset) and a substitution in the first half (which is not very common, as seen in Figure 2, and may point to an injury or other issue). These events were located within a game, and the shots before and after the event were gathered. After that, the xGs of the shots were

compared to see if there was any significant difference in the quality of chances that the team had before and after the high-impact event. This was compared from both sides; for example, when a team has a player sent off, this is evaluated both from the perspective of the man-less team and from the opposing team.

Results

Evaluation of the xG model

While the xG model is not meant to be strictly "correct", it is useful to examine it to see that there were no obvious errors. For this, three metrics were used: the accuracy on cross-validated training data (with 5 folds), the accuracy on test data, and the ROC AUC score. Table 1 shows the metrics and their evaluations. Note that the dataset for shots is extremely unbalanced; 89.35% of the dataset is shots where a goal was not scored, so a classifier where every shot was classified as a non-goal would have this accuracy. No balancing was done for this project, as the dataset should reflect the average shot-to-goal ratio in real games.

As can be seen, the accuracy is quite good at almost 94%, and seems to be consistent across cross-validate training as well as testing data. The AUC score is good as well – an analysis of some xG models (Gelade, 2017) showed that an average model had an AUC score of 0.75-0.82. While the metrics are not necessarily good at showing how well the model performs (due to the choice of features – see the methods section), they show that there are no obvious errors and the model is well-behaved. As a final measure, the total xG of all of the shots in the dataset was calculated. While there are 24 177 goals in the dataset, the total xG is calculated as 24 202 (0.1% error).

metric	value
Accuracy on cross-validated training data	0.9385
Accuracy on test data	0.9380
AUC score on test data	0.9650

Table 1: performance metrics of the xG model.

Analysis of the dataset as a timeseries

As the dataset can be thought of as a timeseries of combined matches (with minutes 1-90 as the time), it was evaluated in this way. Figure 3 shows the mean expected goals for *shot* events (as evaluated by our xG model) over the course of a game. A positive trend can be seen, implying that over the course of the game, the chances to score from shots increases. Note that the number of all events increases over time as well (see Figure 1). A detrended version can be seen in Figure 4. In the detrended version, one outlier can be seen. On minute 46, the chance to score a goal from a shot is 2% less than on average (which is about 11%). This is likely due to the players not being able to create good shooting scenarios in the one minute they have from the start of the second half, though it is interesting to note that this is not the case with minute 1 (though it is negative as well).

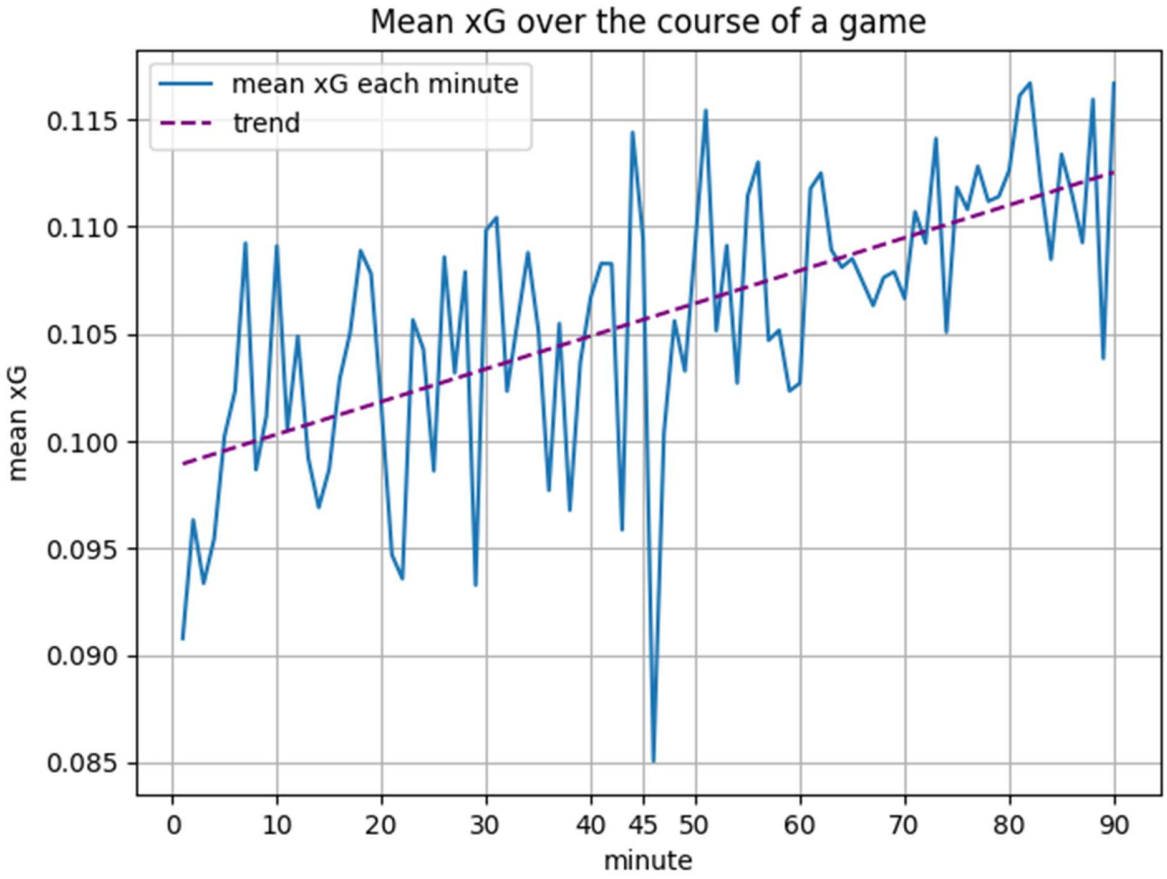

Figure 3: The mean expected goals over the timeseries.

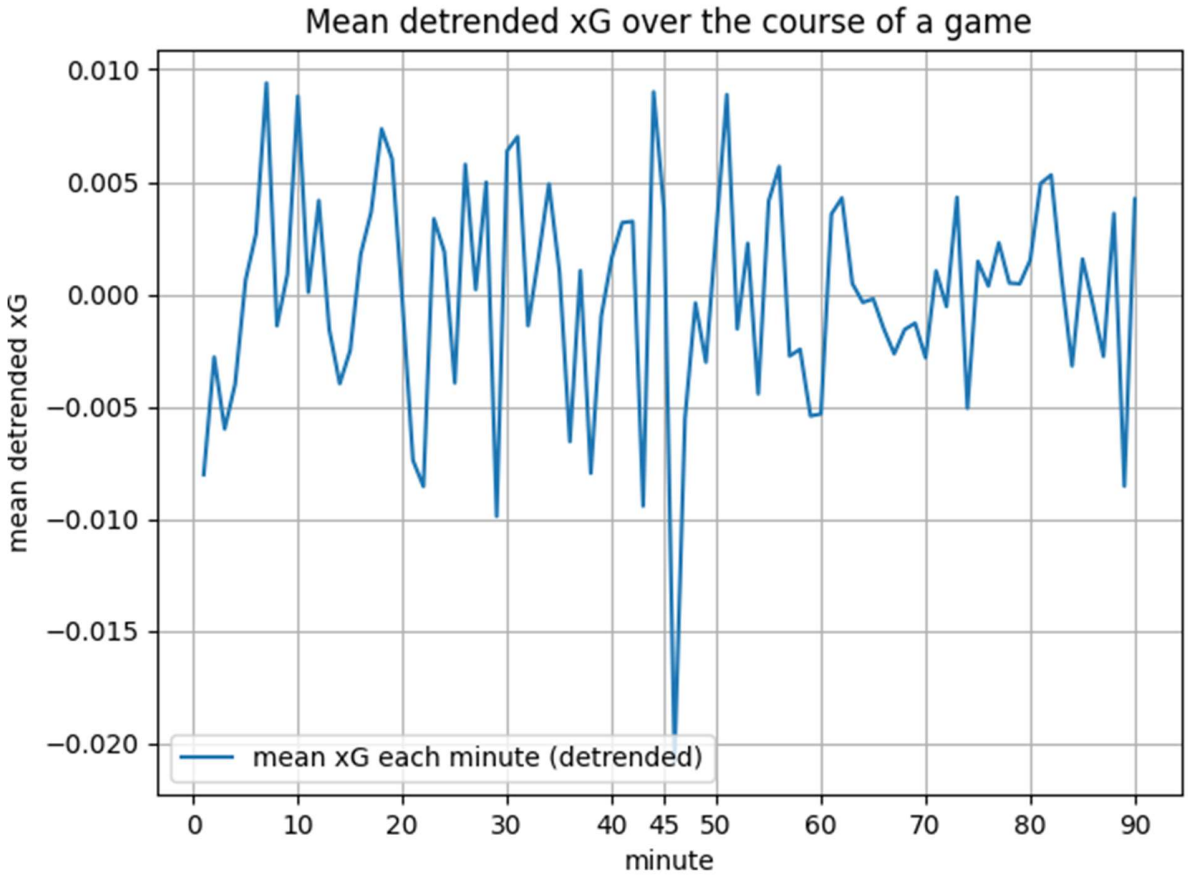

Figure 4: The mean detrended expected goals over the timeseries.

Analysing how specific events influence the rest of the game

A note on standard deviations and detrending

The mean xG values that will be discussed below have two caveats: their distributions have standard deviations (SDs) higher than the means themselves, and they have had detrending applied to them.

With regards to SDs, the samples have means with magnitudes around 0.001-0.01, and they all have SDs around 0.23. This is due to the uneven distribution of the dataset; as seen in Figure 5 and discussed above, most of the shots are not goals, and about 70% of the shots have xGs close to 0. This leads to a non-normal distribution and a high SD, but the data is still appropriate for inter-mean comparisons. Because of these reasons, the SDs are not reported below.

With regards to detrending, because we are comparing two distributions, one of which has shots which happened strictly before the other's shots, it is natural that the later distribution would have a higher mean (due to the time-dependent trend mentioned above). To counteract this, detrending is applied to every xG in the dataset.

The results of the high-impact events analysis can be seen in Table 2. The table should be interpreted as such: the first row should be read as, "How does the quality of chances for a team

change when a player from this team gets sent off?". The **team affected** column denotes which point of view is evaluated; for example, the second row should be read as, "How does the quality of chances for a team change when a player from the *opposing* team gets sent off?". Note that the mean values, as well as the change, are increased 100 times for a clearer view. A t-test comparison was made between the two distributions to evaluate whether they were significantly different. The chosen threshold is the literature-standard alpha of 0.05. It can be seen that all of the distributions, except for the situation where your team misses a penalty, are significantly different. Note also that in this analysis, there could be overlaps; for example, if there were two red cards in a single match, these would be counted as separate matches; of course, the "before" and "after" distributions would be different as well.

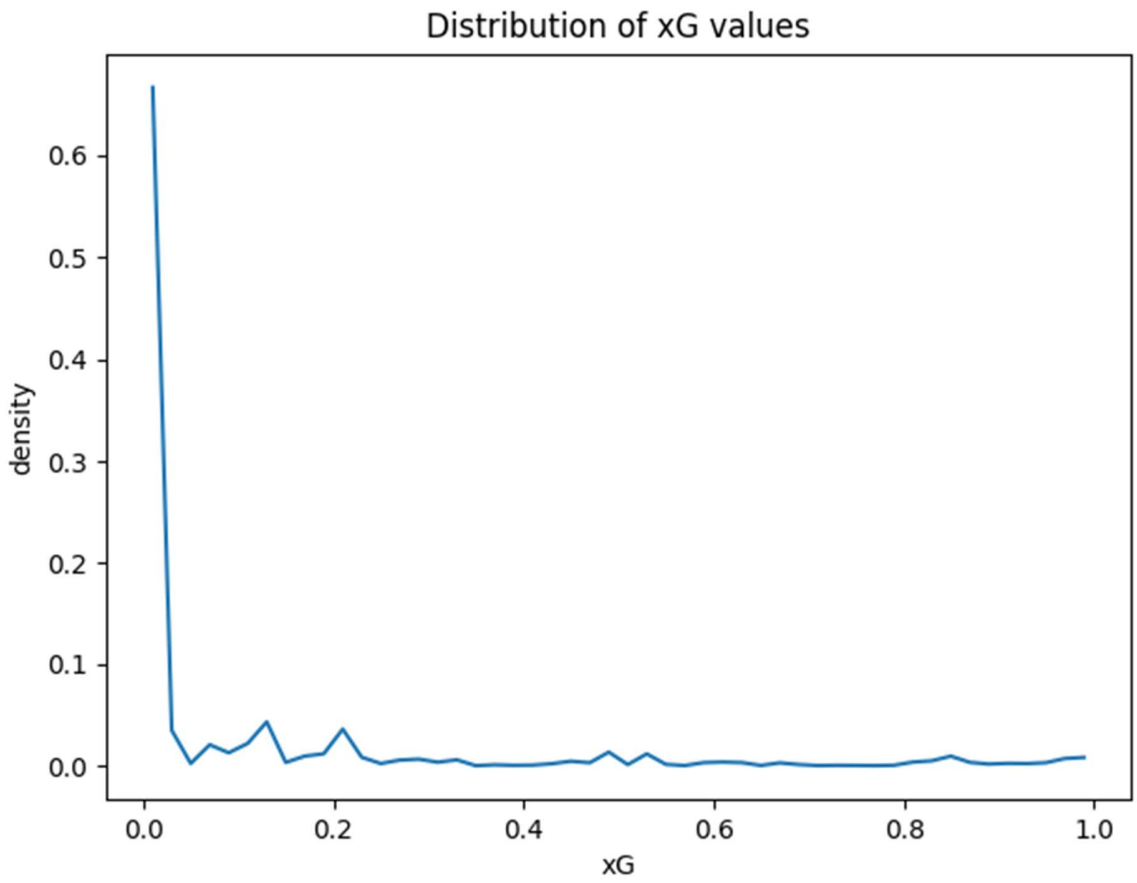

Figure 5: The normalized distribution of xG values for shots in the dataset.

situation	number of datapoints	Team affected	Mean xG before (×100)	Mean xG after (×100)	Change (×100)	p-value of t-test
A player gets sent off	1889	original	-0.132	-1.033	-0.901	0.009
		opponent	1.257	1.983	0.726	0.015
A team gets a penalty	2384	original	-0.386	1.238	1.621	< 0.001
		opponent	0.965	-0.668	-1.633	< 0.001
A team misses a penalty	566	original	-0.168	0.386	0.554	0.286
		opponent	1.375	-0.369	-1.744	0.002
A team scores a penalty	1818	original	-0.449	1.540	1.989	< 0.001
		opponent	0.836	-0.748	-1.584	< 0.001
A team has a substitution in the first half	5945	original	-1.341	-0.178	1.163	< 0.001
		opponent	3.135	0.471	-2.664	< 0.001

Table 2: Changes in the distributions of xGs before and after the high-impact event. The change column is highlighted in green, while the statistically insignificant p-value is highlighted in orange.

Discussion

It can be seen that the method can adequately measure how a single high-impact event will influence the rest of the match. A further analysis of Table 2 can be found below.

The situation where a player gets sent off has reasonable results – the team with fewer players has worse chances for the rest of the game, while the opponent's chances are increased. It is notable that the mean xG was slightly negative even before the sending off happened. This could indicate that it is generally worse-than-average teams that get more players sent off, or that a player often gets sent off due to the decisions they make when the team has a bad performance. It should also be noted that there is a bigger difference in the xG for the sent-off team compared to the increase in xG for the opponent. This could be explained by the fact that often when a team goes down a defender or a midfielder, they swap strategy, such that they have one less striker instead (either through substitutions or by playing the striker in the midfield). That would decrease one's chances of scoring more than one's chances of conceding a goal.

The penalty statistics are slightly skewed. A general trend can be seen, where regardless of whether the penalty was scored or not, the team which got the penalty has an increase in xG, while the opposing team has a decrease. This could be attributed to the fact that penalties are generally awarded for harsh fouls and accompanied by yellow or red cards. Thus, the effect that is assumed to be the effect of the penalty, could instead be the effect of a sending off or a player playing more cautiously due to a warning. However, the pressure a goal creates can be seen from these statistics – while missing a penalty is statistically insignificant for increasing xG, scoring a penalty carries a huge gain of almost 0.02 xG. While analysing goals was not done due to the large number of events, it can be inferred from penalties that a scored goal makes the opposing

team scramble and try to score goals from worse positions, while at the same time playing more offensively and defending less, leading to an increase in xG for the opposing team.

Finally, the substitution statistics are quite interesting. An early substitution leads to a significant increase in xG for one's team, but even more notably, a large decrease in xG for the opponents. Two possible interpretations for this could be: the player substituted played with an injury for some time, leading to the skewed xGs, or the manager had bad starting tactics and made quick tactical changes to counteract what was happening on the pitch. Note that the mean xG values did not change their sign, so even after the substitution, one's team is slightly worse and the opposing team is slightly better, however, there is a definite improvement.

Conclusion

The method applied shows how a single event can be evaluated. While usual xG metrics only take into account features of the exact shot and at most one or two events before the shot, this method can be used to see how the likelihood of good chances can be influenced by events that took place quite a bit beforehand.

Further work could be done in a number of directions. First, a better dataset could always improve and expand on this work; while this data was good, there were some issues, notably the event type *second yellow card* was not properly labelled, with many second yellow cards being labelled as simply *yellow card*. A coordinate system for the location of the event, instead of the current zone method, would likely improve both the xG model and the high-impact event evaluation. A big improvement would be the inclusion of player coordinates for the events. While these have been seen in other datasets, it is not an easy task to collect them, so it is reasonable that this dataset does not have them. However, having coordinates could allow us to raise some other questions: does a foul in a specific place lead to injuries more often? How does the team formation change after a high-impact event? How do substitutions of specific positions influence the game?

Another direction would be a different method of analysis. Currently, the *before* and *after* events are collected into a single distribution, not distinguishing between what team or player made the event, nor in what match they happened. An alternative would be to do just that – for example, take a single team and see how events in this team's matches influence the game. This would likely require a lot more data, as a team generally only plays 38 matches in a season. However, this might fix the fact that during the high-impact events analysed, all of the *before* event means had negative values for the original team and positive values for the opposing team. When analysing only a single team, these could be adjusted.

All in all, this method can be used to evaluate how much a single event influences the rest of the match. This can be used by data analysts to see which outcome is preferable (for example, when a player must choose between giving a good scoring chance for the opponent or making a foul that leads to a penalty) in certain situations. It can also be used to find important factors that could otherwise be left out, such as the early substitutions seen above.

References

Gelade, G. (2017). *Assessing expected goals models. Part 2: Anatomy of a big chance*. Business Analytic. Retrieved May 30, 2022, from http://web.archive.org/web/20190716072916/http://business-analytic.co.uk/blog/assessing-expected-goals-models-part-2-anatomy-of-a-big-chance/

Goodman, M. (2018). *A new way to measure keepers' shot stopping: post-shot expected goals*. StatsBomb. Retrieved May 30, 2022, from https://statsbomb.com/articles/soccer/a-new-way-to-measure-keepers-shot-stopping-post-shot-expected-goals/

Green, S. (2012). *Assessing the performance of Premier League goalscorers*. Stats Perform. Retrieved May 30, 2022, from https://www.statsperform.com/resource/assessing-the-performance-of-premier-league-goalscorers/

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering*, 9(03), 90-95.

Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. WW Norton & Company.

Nandakumar, N., & Jensen, S. T. (2019). Historical perspectives and current directions in hockey analytics. *Annual review of statistics and its application*, *6*, 19-36.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Journal of machine Learning research*, *12*, 2825-2830.

Secareanu, A. (2017). Football Events, Version 1. Retrieved May 30, 2022, from https://www.kaggle.com/datasets/secareanualin/football-events/versions/1

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & Van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, *17*(3), 261-272.

Zuccolotto, P., Manisera, M., & Sandri, M. (2018). Big data analytics for modeling scoring probability in basketball: The effect of shooting under high-pressure conditions. *International journal of sports science & coaching*, *13*(4), 569-589.