Creating a Model for Expected Goals in Football using Qualitative Player Information

Pau Madrero Pardo

Advisor: Javier Fernández (F.C. Barcelona)
Tutor: Marta Arias (Computer Science Department)

Master in Innovation and Research in Informatics
Specialization in Data Science
Facultat d'Informàtica de Barcelona (FIB)
Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

June 29th, 2020

Abstract

The field of sports analytics has been growing a lot in recent years. Sports like baseball and basketball were among the first to embrace it, but football has also taken big steps in that direction. One of the causes is that data analysis allows for the development of new advanced metrics which can provide a competitive advantage.

This project presents a new version of one of these advanced metrics applied to football, the Expected Goals. The metric estimates how likely it is for a shot to end up becoming a goal. We present two different approaches for building the predictors: one that uses player qualitative information and another player agnostic. We then reflect on the importance of the calibration of the probabilities yielded by the models, as well as their possible interpretations, and present some of the applications that can be used to evaluate team and player performance.

We also show the impact each feature has on the models to make their outputs interpretable and to demonstrate that the addition of the player qualitative information is important for the performance of the model.

Glossary

Event Data

Log-like data of football games. It captures information about on-ball events which includes the location and the players involved.

Tracking Data

Positional data of football games. It captures frames consisting of the location of the 22 players and the ball. The frame rate can vary depending on the data provider but it is usually at 25Hz.

OPTA

One of the major football event data providers. The data used in this project came in great part from this provider.

Expected Goals (xG)

It estimates the probabilities of scoring a goal given that a shot is produced.

Expected Possession Value (EPV)

It estimates the value of the possession at a certain moment. It corresponds to the expectation of ending up scoring a goal in that possession or being scored against in the next one.

Open play

It embraces all those situations during a football game where the origin of the attack has not been from a game resuming events such as a corner, free-kick, penalty or throw-in.

Contents

List of Figures

List of Tables

1 Introduction

This project corresponds to a Master's Thesis (TFM) of the Master in Innovation and Research in Informatics (MIRI) on the specialization of Data Science, at the Facultat d'Informàtica de Catalunya (FIB) of the Universitat Politècnica de Catalunya (UPC).

This is a project developed in the UPC with the collaboration of F.C. Barcelona, a multi-sports club considered one of the biggest football clubs in the world, established in the city of Barcelona.

In this project, we will present a new version of an already in use football metric called expected goals. The rest of the introduction is structured as follows. We will start by talking about the topic of sports analytics in general and introducing some of the most important advances there have been. We will do the same with football analytics in particular, and we will mention the different kinds of data available in the market. We will finally list the motivations and goals of the project.

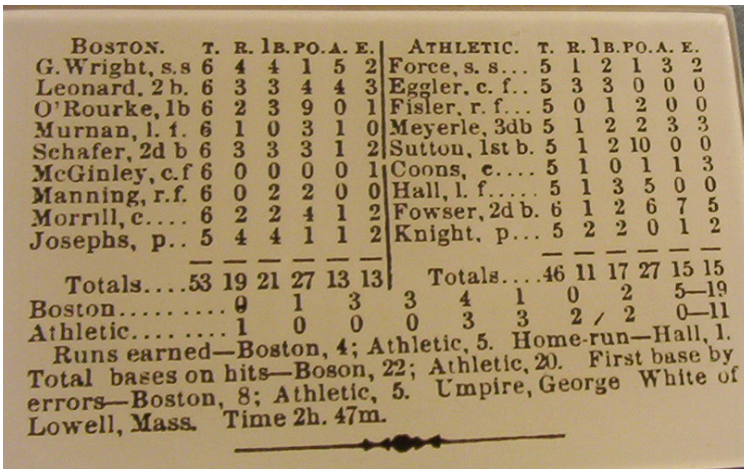
1.1 Contextualisation

The F.C. Barcelona's Sports Analytics department collaboration with this project has been key to the development of this project, especially the support provided by the football coaches and experts.

1.1.1 Sports Analytics

The football data analytics community is relatively young, especially compared to other sports. Baseball was the first sport to have mainstream use of data, as explained by Michael Lewis in his book *Moneyball: The Art of Winning an Unfair Game* [Lewis, 2004], and its subsequent popular film. Michael narrates how in the early 2000s, the Oakland Athletics were capable of being competitive against teams with a much higher budget and to get into the playoffs two years in a row, by taking advantage of analytically proven statistics which were underrated by the old-school baseball community.

The baseball world has been surrounded by statistics almost since the sport was invented. Figure 1 shows a box score of a match played between two American teams in 1876. The very nature of the sport allows for this kind of statistical summaries to be elaborated easily, since every time the game is resumed, it is done from the same static situation — the pitcher throws the ball and the batter tries to hit it.



BOSTON.	T.	R.	1B.	PO.	A.	E.	ATHLETIC.	T.	R.	1B.	PO.	A.	E.
G.Wright, s.s	6	4	4	1	5	2	Force, s. s...	5	1	2	1	3	2
Leonard, 2 b.	6	3	3	4	4	3	Eggler, c. f..	5	3	3	0	0	0
O'Rourke, 1b	6	2	3	9	0	1	Fisler, r. f...	5	0	1	2	0	0
Murnan, l. f.	6	1	0	3	1	0	Meyerle, 3db	5	1	2	2	3	3
Schafer, 2d b	6	3	3	3	1	2	Sutton, 1st b.	5	1	2	10	0	0
McGinley, c.f	6	0	0	0	0	1	Coons, c....	5	1	0	1	1	3
Manning, r.f.	6	0	2	2	0	0	Hall, l. f.....	5	1	3	5	0	0
Morrill, c....	6	2	2	4	1	2	Fowser, 2d b.	6	1	2	6	7	5
Josephs, p..	5	4	4	1	1	2	Knight, p...	5	2	2	0	1	2
Totals....	53	19	21	27	13	13	Totals....	46	11	17	27	15	15

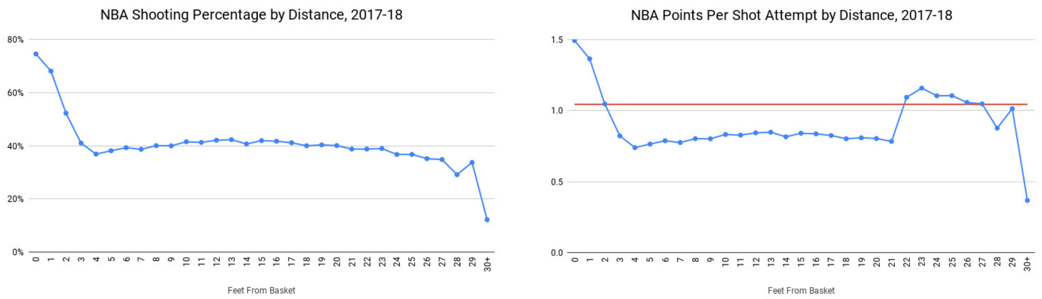
Boston..........	0	1	3	3	4	1	0	2	5—19
Athletic.........	1	0	0	0	3	3	2 /	2	0—11

Runs earned—Boston, 4; Athletic, 5. Home-run—Hall, 1.
Total bases on hits—Boson, 22; Athletic, 20. First base by
errors—Boston, 8; Athletic, 5. Umpire, George White of
Lowell, Mass. Time 2h. 47m.

Figure 1: Baseball box score from a match between Boston and Athletic from 1876 [Blog, 2020]

But having access to all these stats is of very little use if you do not know how to correctly interpret them. And that is what the front office of the Oakland A's did to gain an advantage against their competitors at a lower cost. They proved analytically that metrics such as *on-base percentage* and *slugging percentage* were much better indicators for offensive success than the metrics that were being used by most teams, and they used them to find undervalued players in the market. Since then, most teams in the MLB have incorporated data analytics to help to scout new talent and making in-game decisions.

Another example of the impact of data analytics in sports is basketball. It started a few years later than in baseball, but the progression has been significant. To the point that nowadays [Goldsberry, 2019], almost all the teams in the NBA have a data analytics department. One of the biggest contributions has been what is popularly know as *the three-point revolution.* Data helped them realise that shooting from



(a) Percentage of shots converted to goals by distance

(b) Same as (a) multiplied by the points obtained at that distance

Figure 2: Basketball shooting data of the NBA's 17/18 season. Data from NBA Savant. Source: The Ringer [Kram, 2019]

mid-range wasn't worth it, given that shooting from a little farther away could get them 50% more points.

This phenomenon is clearly illustrated in Figure 2. Plot (a) shows how making a shot from right next to the basket will most likely get you more points, but from 3 feet upwards the chances are more or less similar until you get to more than 30 feet. If we now take a look at plot (b), we will see how many points we would expect to get at each distance, given the probability of scoring and the number of points awarded for a goal at that distance. It is clear to see that three-point shots have a much higher value of expected points than two-point shots from mid-range.

This realisation led to many teams changing their attacking strategies and changed the way the game was being played from that moment onward. One of the main promoters of this change and the basketball data analytics is Daryl Morey, the General Manager of the Houston Rockets. Ever since he joined the team, he has tried to build a team and develop strategies which allowed to take advantage of this discovery. Morey now hosts one of the most popular data analytics conferences (the MIT Sloan Sports Conference).

Figure 3 shows how the metric *3PAr* — 3-Point Attempts ratio — has evolved in the recent years of the NBA. This metric indicates which percentage of the total shots of a team in a game have been three-pointers. The plot illustrates how many games per season a team got a *3PAr* greater than or equal to 50% for the last 40 seasons.

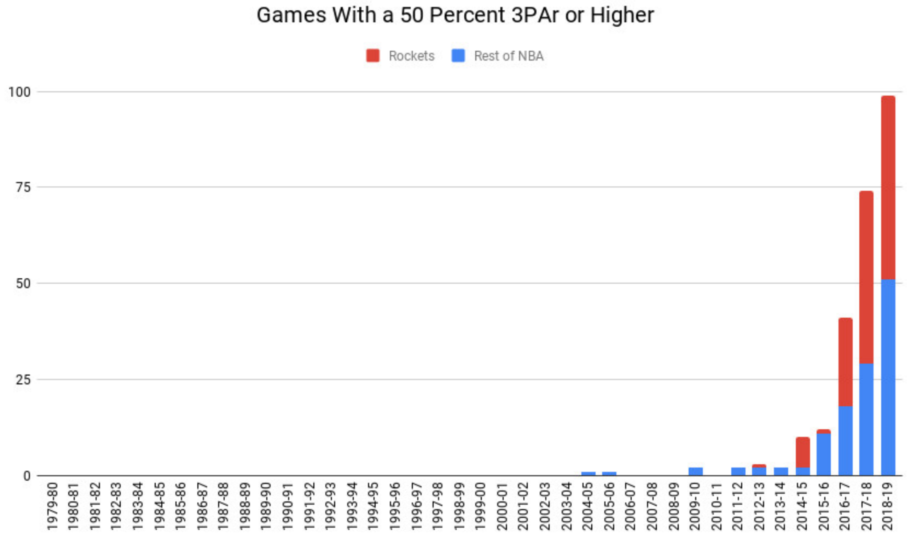


Figure 3: Evolution of the number of games with more three-point shot attempts than two-point in the NBA. Source: The Ringer [Kram, 2019]

We observe that there are games of this kind only in the last 15 years and that in the last 5 years the number of games has increased a lot. The plot also allows us to depict that Morey's team has contributed with as many games to the total as the rest of the NBA teams combined.

1.1.2 Football Analytics

Football is still far behind these two sports on the matters of sports analytics advancements, but every year more and more teams are starting to invest money and resources to it. There have already been important developments though, such as the creation of the metric Expected Goals (xG)[Lucey et al., 2014]. The xG of a shot corresponds to the probability that this shot becomes a goal. So it allows us to have a qualitative measure of danger to the opponent's goal, and not only a quantitative one — as would be for example the number of shots.

Hopefully, it will just be a matter of time before football catches up with sports analytics leaders such as baseball and basketball. A further discussion on the most important football analytics developments published until now can be found in Chapter 2.

Football Data

One of the reasons why football is behind basketball and baseball on sports analytics is because it is much harder to get access to data. Although some data providers are starting to release public datasets [Statsbomb, 2020] [Sports, 2020b] to try to draw more people into working on this field, it is still expensive to get access to large and complete datasets. Currently, there are two main types of football data.

The first kind is known as event data. This type of data consists of detailed information about events that happen around the ball during a match, such as passes, shots or fouls. So it can be seen as a log of all the on-ball actions that happen during a game. The information registered of each event normally involves the location where it happened, the players involved, the instant in the game when it happened, and sometimes other event-specific attributes. Event data provider companies normally manually tag these events using the TV broadcast video of the games.

The second kind of data is what is known as positional or tracking data. The tracking data includes the XY locations of all the players on the pitch and the ball, during the whole game. An example of the visual representation of an instant of tracking data can be seen in Figure 4. The positional data can be recorded using different techniques, the most precise one being the use of GPS technology. The drawback of this way of obtaining the data is that all players must wear a device capable of tracking their locations, which must be compatible with a system of sensors installed on the stadium where the match is being played. And to get the locations of the ball, it needs a chip embedded inside. Since there are still no agreements between teams and championships to standardise all these procedures and systems, the GPS approach remains unfeasible.

The most popular method to obtain tracking data for the moment is optical tracking. It consists of obtaining the positional data directly from the video feed, by making use of computer vision technologies. It can be extracted from any kind of the available match video types: panoramic video, tactical video and TV broadcast video. The first kind is rarely used since it is quite uncommon to record games using static cameras that allow capturing the whole pitch. The tactical video is normally used by football analysts to analyse previous games, so it is fairly popular, and therefore available. In this kind of video, the image captures a wider view space than the TV broadcast. This is helpful for both analysts and optical tracking algorithms since it allows them

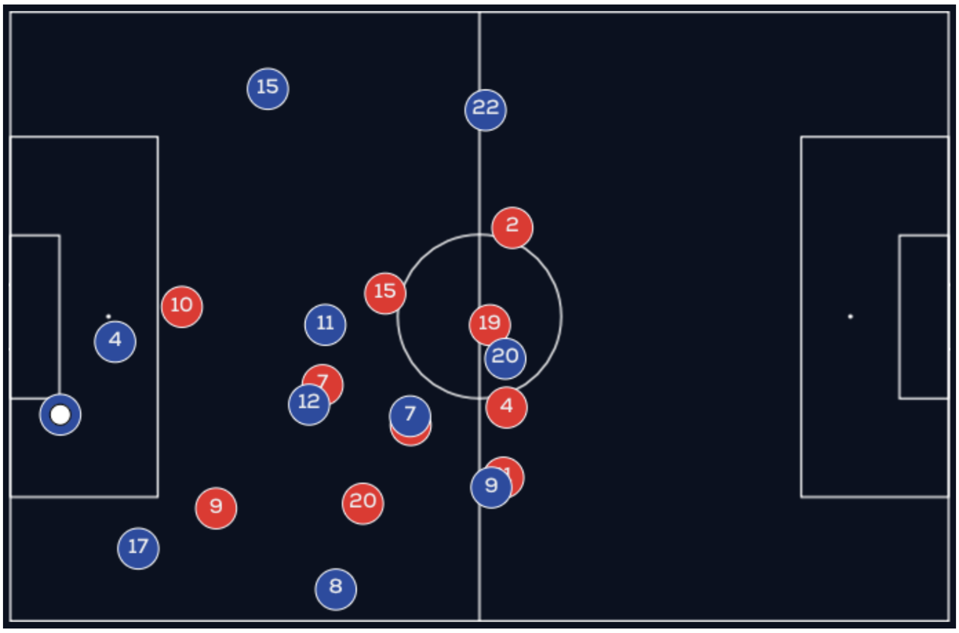


Figure 4: Tracking data for an instance of a game plotted over a 2D football pitch. Each coloured circle marks the location of a player, with the number being the players' squad number and the colour denoting the team they belong to. In this frame, the blue team's goalkeeper has the ball. Source: Metrica Sports [Sports, 2020a]

to see the locations of players not so close to the ball, which may not be possible with the TV broadcast video because it tends to be zoomed in to where the ball is. Another drawback of the TV broadcast is that replays may sometimes cause some small parts of the game to be missed, but the main advantage is that is the most widely available kind of video by far.

As it happens with the dilemma of which kind of positional data to use, or which type of video to extract optical tracking from, when deciding between using tracking or event data, the answer comes to trade between availability and quality. Since getting event data is normally much easier than getting positional data when positional data is available, it is normally accompanied by the correspondent event data as well. This combination is currently the most complete data mix you can get from a game, at least from a tactical data analysis perspective. But event data by itself can also give very valuable information, and it is much cheaper to get. So if you need a very big dataset and have limited resources, it may be the only achievable option.

1.2 Motivation

Developing new advanced football metrics is key to measure team and player performance. Basic metrics have been used in the past in football to try to explain what happened in a match, or at least part of the performance of a team or a player. Some of these metrics, such as percentage of time in possession of the ball, number of shots or distance covered by the players are still being used, especially by the sports media. Part of them can be misleading when looked at individually. Take for example the number of shots. You could wrongly assume that a team with a much higher amount of shots would dominate offensively an opponent with fewer shots. If most of the shots from the first team were from afar whereas the opponents' shots were clear goal chances, the odds of scoring more goals would probably be with the opponent.

The development of more advanced metrics can help give more precise insights into the data. Following up on the previous example of the number of shots, if we took a look at the xG values of the shots taken by both teams, we would easily see that the first team was much less likely to have scored a goal than the second one.

That is why this project aims to develop an xG model, to provide with a more precise and accurate way to assess on the own team, opponent and player performance analysis. In addition to that, we want to include into the model information related to the quality of the shooting player, to have a more accurate prediction on team performance, as well as being able to have a more tailored performance metric to each player.

1.3 Goals

The main goals of this project that will have to be fulfilled are the following:

- **Build a basic xG model**
  Building a first model without the shooter information will help to have a baseline to compare the final model to.

- **Build a second xG model with qualitative shooter information**
  This second model will be the one we will use from now on to compute the xG of every shot.

- **Calibrate the models**

One of the most important steps will be to check that both models are calibrated. If they are not, their outputs will not be interpretable as probabilities.

1.4 Master thesis outline

After this introduction, the rest of the document will be structured as follows:

- **Chapter 2: Related Work** dives deep into the main studies and developments made on the field of sports analytics, as well as the existing publications on the building of xG models.

- **Chapter 3: Methodology** describes the process followed starting with the data acquisition and finishing with the evaluation of the models.

- **Chapter 4: Experiments** presents the results obtained from the models and shows the applications they can have.

- **Chapter 5: Conclusions** checks whether the initial goals have been achieved. It also summarises the findings uncovered during the realisation of the project.

2 Related Work

In this section, we will go through the previous work done on sports analytics. We will centre our attention on football analytics, but we will mention research on other sports when they were the source of inspiration for their football counterpart.

In football, as in many other sports, there are three main fields of research: sports science, medical and tactical. We will focus on the latter.

In this project, we want to estimate the probability of a shot attempt becoming a goal. That can also be seen as a way of valuing the player's efforts on that action since it will give us a sense of the difficulty of the shot and we will be able to compare it with the final result — goal or no goal. So when looking for previous research on tactical football analytics, we will pay extra attention to studies focused on action evaluation.

2.1 Valuing Player Actions

The first stat we will talk about is called *plus-minus* (+/-). It was first introduced in ice hockey and basketball, but it has recently been adapted to football as well. An advanced version of the statistic was presented by [Sill, 2010] on basketball whereas the football version was published by [Kharrat et al., 2017]. The +/- metric consists of measuring the contribution a player has on the team's success. A simpler approach is computing the difference in points/goals scored by both teams while the player was playing. That can work on basketball given the elevated number of points per game and allowed substitutions. The work in [Sill, 2010] takes into account who the player played with and against. The author performs a regression on game snippets with no substitutions to produce a rating for each player given the outcome of the snippets, the outcome of the game and the participation of each player on which snippets.

The authors in [Kharrat et al., 2017] defend that the previous approach cannot be applied in football given the low scoring nature of the game. They present an alternative which uses the expected goals of the teams' shots as a measure of offensive production instead of just the goals, and an estimation of expected points to determine the outcome of the game. They also apply a regression afterwards to determine the players' ratings.

The two publications referred to until now talk about measuring the value added by a player in a game, but just considering the performance of the team on their presence. [Power et al., 2017] go one step further and evaluate the players' performance by analysing how they perform the most repeated event in a football game: the pass. They do so by using supervised learning approaches on tracking data to predict both the risk and the reward of a pass. The risk responds to the difficulty of the pass, or how likely it is to become unsuccessful, whereas the reward represents the probability of a shot occurring within 10 seconds after the pass. The authors proceed to present several advanced metrics based on these two concepts such as *Passing Plus/Minus* — which evaluates the success of a players passes taking their risk into account — and to define the concepts of *Difficult Pass* and *Dangerous Pass* based on the distribution of the passes' risk and reward correspondingly.

A new framework which allows for the evaluation of players' actions was presented by [Cervone et al., 2014] on basketball. Their idea was to assign a value to every moment of the possession of a team based on the number of points they are expected to score and called it *Expected Possession Value* (EPV). The authors created a possession model which allowed them to estimate the probability that the player with the ball made each one of his available decisions, as well as the resulting EPV after making each one of them. That is key for player performance analysis, given that it allows for the evaluation of player decision-making, and see how often they choose the option which increases EPV the most. Following this line of thought, the authors also present the *EPV-Added* metric, which quantifies the value added to the possession every time a player acts with the ball. The metric is computed by simply adding the difference between the EPV at the end of an action and the EPV at the start of it for every action carried out.

[Fernández et al., 2019] recently published a similar approach to [Cervone et al., 2014] but applied to football. The authors also present an *expected possession value* framework but with the necessary adjustments to account for the differences between both sports. That includes the looser notions of possession, the possibility of passing to open spaces — as with through balls — or the lack of time limit for possessions. Their model uses spatiotemporal tracking data to produce a prediction of the outcome of the possession encoded as a real number in the [-1, 1] range. Values close to 1 represent a high probability of a goal by the attacking team, whereas values closer to -1 indicate a likelihood of an immediate possession change and a posterior goal by the opponent. As in [Cervone et al., 2014], they predict the likelihood of every possible decision a player can make while in possession of the ball, and then the expected value of the possession after executing that action. They decompose the possible actions a player can perform into keep driving the ball, passing to each one of his teammates and shooting. Note that the model that predicts the expected value after a shot corresponds to an expected goals model. For each possible action, they also account for the danger the opponent can produce if they recover the possession of the ball, weighted by the likelihood of them doing so. Just like [Cervone et al., 2014], they use the framework to value player actions with the *EPV-Added* stat.

More applications derived from the framework in [Fernández et al., 2019] are presented by [Llana et al., 2020]. They use both the predicted increase in the EPV for potential passes and the actual increase of passes to measure the quality of offensive and defensive off-ball actions. They present the concept of *off-ball advantage* to identify moments where players of the attacking team who do not have the ball find themselves in a location where they can receive the ball and increase significantly the value of the possession. The authors also introduce an advanced method to assign a responsibility area to each player of the defending team to identify the commonly used concept in football of "passes behind the back". They use the danger of these passes determined by their *EPV-Added* to assign defensive responsibility qualitatively. Finally, they introduce a new metric called *Effective Value Added* which distributes the offensive and defensive production among the different players using the *EPV-Added* values of their actions.

[Link et al., 2016] take a different approach to the concept of possession value and describe it as *Dangerousity.* Therefore, the player's *Action Value* is determined by the degree in which he made a situation more dangerous after he no longer possesses the ball. The level of *Dangerousity* is determined using four factors: the *Zone* or location of the ball, the degree of *Ball Control*, the level of *Pressure* put on the player by the defenders and the *Density* of opponent players in front of the goal. Although they use tracking data to compute most of these factors, the criteria used to determine their actual values was established by football experts after analysing a large number of match situations, instead of using automated learning methods.

Other EPV-like approaches built using event data have been introduced by [Gyarmati and Stanojevic, 2016], as well as [Decroos et al., 2017], which was extended by [Decroos et al., 2019] and finally by [Bransen and Van Haaren, 2020]. In [Gyarmati and Stanojevic, 2016], they focus on the value-added by the players' passes — as in [Power et al., 2017] — by partitioning the field into squares and assigning them values. They determine the value-added of each pass by calculating the difference between the values of the destination zone of the pass and its origin zone. The authors of [Decroos et al., 2017] determine player action values by assigning a percentage of the rating of an attacking phase among its actions using an exponential-decay-based weight system starting by the end of the phase. They determine the rating for an attacking phase as the proportion of similar phases that end in a goal. In [Decroos et al., 2019] they present a second approach which predicts, using supervised learning methods, how likely is a certain player action in a certain game state to produce a goal in the following 10 or fewer actions. And similar to [Cervone et al., 2014] and [Fernández et al., 2019], the action value is determined by the change in this predicted probability before and after the action. Finally, they evaluate this change on sequences of actions between different pairs players in [Bransen and Van Haaren, 2020] to assess the chemistry between members of the same team.

2.2 Academic Work on Expected Goals

Expected goals became one of the most popular metrics a few years back in the football analytics community. And so induced a lot of football fans and data enthusiasts to share their knowledge and researches, especially on xG, using less rigorous publication sites such as web blogs. Although some of these blogs have published valuable investigations, we will focus on the academic literature.

To our best knowledge, the first academic publication on expected goals was done by [Lucey et al., 2014]. They build a single logistic regression model to predict the likelihood of a shot based on features which they group into location, context, defending and attacking. Most of the used features are extracted using a mix of event and tracking data, such as the defender proximity and formation or the attacking team movement relative to the opposition.

A similar solution to the problem is presented by [Eggels et al., 2016]. They also use event and tracking data to define the model's features, but they try four different classifiers — three tree-based classifiers and logistic regression — instead of going just with logistic regression. The authors additionally include as features the shooter's and goalkeeper's quality obtained from EA Sports FIFA video-game publicly available database.

The published work most related to the case study of this project is probably found in [Kharrat et al., 2017]. That is so given that they are also limited to the use of event data to build the xG model. But as mentioned before, the ultimate goal in [Kharrat et al., 2017] is different from the ones of this project, considering that the authors pretend to use the model as a target for a plus-minus player rating. That is why they decided not to include the shooting player's ability as a model feature, provided that it would induce a feedback loop. Instead of building just one generic model to predict the outcome of any shot, they build four specialized models to use depending on the kind of shot: free kick, header, open play or penalty. They try on different classifiers for each kind of shot and select the one that yields better results for each one of them. The model features are more basic than in [Lucey et al., 2014] and [Eggels et al., 2016] due to the lack of positional data. They are mainly derived from the location of the shot and the goalkeeping skills obtained from the FIFA video-game as well.

Two additional papers should be mentioned when conferring about this topic since they present different approaches not seen in the previously mentioned work. The first one was presented by [Rathke, 2017] and shows a simplistic solution which consists of dividing the shots into 8 studied partitions on the field and defining their probability as the percentage of scored shots on their zone. Finally, the research from [Spearman, 2018] takes a step further on the paradigm of the expected goal to evaluate off-ball chance creations as the positioning of players in scoring locations, even if the ball is never delivered to them.

3 Methodology

This section thoroughly describes the process followed during the development of the project with the final goal of building two xG models able to predict the probability of scoring a goal given a shot. It is structured as follows:

- **Data Acquisition** details where the data used for building the models came from and what is its format and structure. Then explains how it will be stored to perform data exploration and data enrichment.
- **Data Preparation** narrates the processes followed to get the data ready before building the dataset, byways of formatting and enriching the data.
- **Dataset** explains the final considerations taken to produce the dataset that will be used to train the models.
- **Model Design** formally defines the problem we want to solve.
- **Model Implementation** lists the classifiers and metrics used to build the models.
- **Model Results** shows the results obtained from the classifiers and reflects on them.

3.1 Data Acquisition

In the contextualisation section, we discussed the different types of data that are currently available. Given the two different types of football data available, two different generic kinds of models can be built. We can either use tracking data to include features related to the position of the 22 players or use event data to take advantage of its abundance. The models using tracking data will not have abundant shots, given that this kind of data is harder to get. So for this project, we will choose to work with event data. Table 1 shows the availability of event data at the moment of the development of the project.

	Games	Shots	Goals
Event Data	4888	118211	12309

Table 1: Number of games, shots and goals available.

3.1.1 OPTA Data

OPTA Sports is one of the leading football event data provider companies in the world. They cover over 1.000 leagues and competitions worldwide, and they provide detailed tags. For this project we had access to the OPTA data of 6 different competitions: *Champions League* (Europe), *Bundesliga* (Germany), *Serie A* (Italy), *La Liga Santander* (Spain), *La Liga SmartBank* (Spain) and *Copa del Rey* (Spain). This data corresponds to the seasons 2017/18, 2018/19 and the games of the 2019/20 played until March.

3.1.2 FIFA® video-game Data

Apart from the event data, we will need complementary qualitative information about the players to build the second xG model. This information needs to be individualised per player and to account for the player's level of quality, especially for those traits related to the shooting.

FIFA is a football video game, made by the company Electronic Arts (EA). They have a complex system that rates each player ability, to set differences in the levels of manageability of the players you can control in the game. They release a new version of the game every year, so they keep the rates updated to the current performance level of the player. Given that there is currently no provider that produces data at the level of detail they require and for the number of teams and competitions they present in the game, their rating system is based on the opinion of thousands of volunteer coaches, scouts and football fans whose local knowledge helps avoid inconsistencies in the database and improve its accuracy[Murphy, 2019]. All this data is then overseen by a team in EA, who is responsible to build the player database from it.

This dataset is from a public source, as EA publishes all the player rating data on their website [Sports,]. We have downloaded the dataset through a Kaggle project [Leone, 2019] that scraped that website to obtain the player ratings from the games FIFA 15 to FIFA 20. This dataset is structured in different *CSV* files, and on the Data Preparation section, we will go through how we can integrate it with the OPTA data.

3.1.3 Preparation of the environment

The data comes structured as XML files. We will need to store this data into our database in a structured manner to ease the processes of data exploration and data enrichment. We have chosen to use a document-oriented database, in specific MongoDB, due to the flexibility it allows on the structure of the data. There is still no common standard to structure the event data supplied by the providers, so this flexibility helps when integrating data with different structures.

The first step will be to store the XML structured data into our MongoDB database. The OPTA data is accessed through different feeds, that provide different kinds of information. For example, the feed *F40* contains data on competitions, teams and players. We will create a new collection on our database for each of these entities and store there the information provided by this feed. We will do the same with the match information from the feed *F01* and also with the event data from the feed *F24*.

The only significant change we will need to do on the structure of the data to store it into our document-oriented database is the event data itself. On the XML, some of the events have a list of qualifiers to enrich the information of the event. The passes made with the head are an example. The mentioned list will contain an element which will identify the pass as a header. We have seen fit to convert this qualifier list into new boolean parameters in the database document which indicate whether each qualifier appears in the event, or not. This new structure is more intuitive and will yield faster access times as accessing a boolean field of a document is faster than checking if a value is contained in a list field.

This new database allows us to have a data environment with easy access to explore it and to modify it to improve and enrich the information it already contains.

3.2 Data Preparation

Once we had the raw data saved in our database, the following step was to explore it and process it to have the data ready for building the dataset. We consulted football experts from FC Barcelona to understand which contextual situations may improve the chances of scoring a goal after a shot.

3.2.1 Creation of new concepts

Football experts use dozens of concepts to analyse and describe the game. We decided to base the algorithms that define these concepts on rules as well because it is the most direct way to translate the experts' knowledge into data. The experts also helped us set those rules. Another advantage of choosing this approach is that knowing that the foundations of the models are based on their way of thinking, will help the analysts to better understand and interpret the results obtained.

Figure 5 presents the flow followed for the creation of each new concept on the database. The diagram shows that it is an iterative flow, considering that some complex concepts will be hard to define with rules at the first attempt. Note that the underlined steps of the process indicate that those tasks are performed alongside a football expert. We will start by a generic definition of the flow, and in section 3.2.2 we will see the specific procedure implemented to create the concept of an attack.

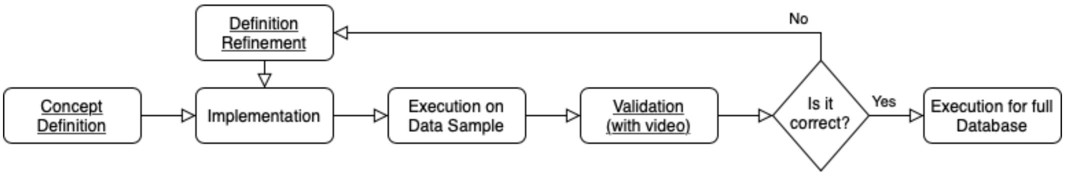


Figure 5: Iterative flow of creation of new concepts. The underlined tasks are performed along with a football analyst

The process starts with the first definition of the rules needed to identify the concept in the data. Having two profiles with two different mindsets helps a lot when translating the game concepts into rule-based algorithms. The football analyst brings all the knowledge about the game and experience in all its aspects, whereas the data analyst knows what can and can't be extracted from the data and in which ways. The first step normally starts with an initial definition by the football expert of how they understand the concept and identify it when watching a game. Then the data expert makes a first suggestion of how that could be translated to identify the concept in the data, and the discussion proceeds until a compromise is reached.

Once the initial definition of the concept is agreed upon, it is time for the second step: the implementation of the algorithm capable of adding this new concept to the database. The complexity of the implementation tends to be very related to the complexity of the concept definition. So for the first version, we try to look for simpler approaches and make some adjustments later on if necessary. This will help a lot with the interpretability of the decisions made by the algorithm.

The next step after the development of the implementation will be to test it out on a sample of data. Trying the new code directly on all the matches in the database may sometimes take a lot of time, or even provoke changes to the database difficult to revert in case of an error in the implementation. So we will first try it on a subset of games. If the new concept occurs often in a game, using one or two games will probably be enough to cover all cases, whereas a less common concept will probably need more.

When validating that the concepts identified by the algorithm are correct, the help of the football analyst is needed again. The main tool the analysts and coaches use to transmit their ideas and arguments is the video. So being able to link all the concepts to video will be important since they will be the ones receiving our insights, as well as the ones who will help us with the sport-related doubts we may have. The validation will, therefore, be performed by checking that the fraction of the game's video that the algorithm has identified as the concept, actually corresponds to the concept, according to the expert. In the example of the concept of an attack, we will have to validate that both the start and the end of the attack are correctly identified and that there were not any opponent attacks in the middle. This validation procedure will allow us to identify false positives, and figure out which parts of the definition of the concept need to be refined. If we consider that our definition is not generic enough and may miss some cases — true negatives —, the football analyst will manually identify the concepts in the games of the data sample and check them against the ones automatically identified to see if there are missing examples.

If misidentifications are detected during the validation, the initial definition of the concept is modified to include the necessary changes and consider the special cases that were not considered. This step is also done with the collaboration of both experts, to ensure that the new strategy is feasible and conceptually correct. The changes are then applied to the implementation, and this process along with the execution using the data sample and validation is repeated until all the examples of the concept are correctly identified. The final step is to execute the new algorithm for the games in the database.

3.2.2 Contextualising the shots

The process of describing the context of the shots using data begins by creating very basic concepts that will, later on, allow us to create more complex ones built upon them. The concept of an attack or possession can serve as an example. A simple schematic of the procedure can be observed in Figure 6.

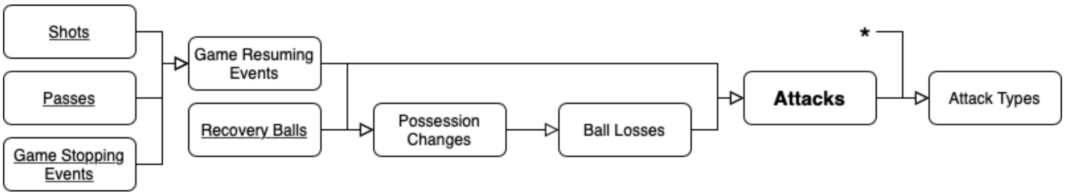


Figure 6: Flow for the attack creation. The events underlined are tagged by OPTA.

The most basic concept we will need to build the attack concept is the game resuming event. This concept will group all the possible events that resume the game after being stopped, such as goal kicks, throw-ins, or kick-offs among others. To do that we will create a new instance of the concept on the pass or shot which came right after the game was paused. Luckily, the passes and the events that pause the game are provided by OPTA, so we won't need to create them.

Another event tagged by OPTA is the ball recovery, which we will use together with the game resuming event to create the concept of a possession change. This one will be as simple as creating a possession change instance every time either a game resume or a ball recovery happens. In cases such as fouls, where the team with the possession before the foul keeps the possession afterwards, we will create two changes — one to the team who committed the foul and the second one to the team who will serve the free-kick — because it will indicate that the first attack has finished and a new one will start afterwards.

Now that we have the concept of possession change we can define the ball losses since we can identify every time a player lost the ball to an opponent — the opponent recovered it — and when the player threw the ball off the pitch. The ball losses along with the game resuming events will be enough to create the attack concept. The game resuming events will always determine the start of a new attack, and the ball losses the end of it. In case of a ball loss to an opponent, it will mark both the end of an attack and the start of a new one.

The process of the attack creation was not as straightforward as it may seem, because of the special cases the football analysts set. For example, they considered that a new attack could not start unless the team in possession had the ball fully controlled. So when considering whether the previous attack had to end and start a new one, we had to check first if the new team kept the ball for a certain amount of time or were able to perform a successful pass or shot event.

The last part of the flow in Figure 6 shows that we use the attacks along with other concepts to finally construct the attack types. These will be used to define part of the context of the shots like for instance determining whether they belong to a play after a corner kick, free-kick, throw-in, penalty, or just regular play. This was not an easy task either, provided that the football experts consider that after each one of these previously mentioned events, there can be a transition to regular play inside the same attack. So establishing the changes between attack types was necessary to correctly define the attack type a shot belonged to.

The final contextual information we added to the shots is the one related to the previous pass that leads to the shot. OPTA adds certain information to passes like tagging them as crosses or through balls. After a careful review with the football analysts, we realised that some specific passes they considered as crosses were not being marked as such, so we had to develop the cross concept to identify the missed cases by OPTA.

All the contextual information of the shots was condensed into a single variable, for we considered that all the categories were exclusive from each other, or at least we wanted the model to consider them as such. The final list of categories is the following:

- **Corner**: Shot coming from a corner kick. As explained before, it can be either a shot directly from the cross or from the derived play, but limited to until the attack type changes back to regular play.
- **Direct free-kick**: Shot from a direct free-kick.
- **Indirect free-kick**: Shot coming from an indirect free-kick. As in corners, it can be any shot until the attack type changes.

- **Outside penalty box**: Any shot in regular play shot from outside the penalty box.
- **Without previous pass**: Shot during regular play where the shooter started the attack by recovering the ball.
- **After rebound**: Shot that happened right after a previous shot.
- **Second play**: Shot after an opponent touched the ball on a pass from a teammate.
- **Cross**: Shot in regular play after a cross. Typically a cross is a pass from one of the lanes on the last third of the field, to somewhere in the centre of the penalty box.
- **Through ball**: Shot in regular place after a through ball. A through ball is a vertical pass that overcomes the last line of opponent defenders.
- **Back pass**: Shot in regular place after a pass coming from near the end line and going to a more backward and centred location.
- **Pass from penalty box**: Shot in regular play after a pass coming from inside the penalty box.
- **Pass to penalty box**: Shot in regular play after a pass coming from outside the penalty box.
- **Drives to penalty box**: Shot during regular play where the shooter is the one who gets into the penalty box by driving the ball.

3.2.3 Integration with the FIFA® dataset

The next information needed by the dataset after the context is the player individual information. As mentioned before, we will use public data from the FIFA video-game. The goal of the integration is matching each one of the players who made a shot in our database with data from OPTA with the corresponding player in the FIFA database. OPTA and FIFA use different identifiers on their database, so we will need to match them using player information available on both databases.

The most common way to identify players in real life is of course by their name. Players rarely use their full name on their professional career, but most likely just their first name or last name, or sometimes a nickname that may or may not be related to their real name. So both databases have several columns related to the player's name, but none that exactly matches the players in both databases. Our goal will be to compute the string similarity between these name columns from both datasets to determine whether they are referring to the same player or not. We found that the measure that worked best for matching names was a string sequence similarity ratio first presented in [Ratcliff and Metzener, 1988]. Its main idea is to find the longest contiguous matching subsequence and then do the same thing recursively with the pieces of the sequences to the left and right of the first one. The similarity of the evaluated is then normalized by their length, so the metric becomes a ratio. That was key for our problem, given that it did not penalize as many names of different lengths — for example when one of the databases used the full name and the other just the last name. Other metrics such as the Levenshtein distance [Levenshtein, 1966] were also tested, which refers to the minimum number of character edits required to change one sequence into the other. We had to invert it though provided that being a distance it is a measure of dissimilarity instead of similarity. But it did not provide results as good as the previously mentioned ratio. So from now on, when we speak about string or name similarity, we will be referring to the string sequence similarity ratio.

Due to the huge amount of players in both databases, computing the similarity between each possible pair will be very time costly. Furthermore, it will yield to wrong matches, given that with so many players some of them will have the same name. Therefore we will have to try a smarter approach.

Our first attempt was to group them by team and season before computing the similarities. The FIFA dataset contains the attributes of each player for every season they have played in one of the available competitions, so the only thing we were missing was the team which the player played for that year. We scraped that information from the original website [Sports,], and matched the teams with the ones from the OPTA database using similarity between their names — grouping them by competition first. Only after computing the player name similarities between the players of the same team and season, we realised that players sometimes change teams mid-season, which was contemplated by OPTA but not FIFA. That caused for all these players to be mismatched.

The approach we finally decided to use was to group the players by their date of birth. This information was also available on both databases, so that helped a lot the process, plus it improved significantly the computation time of the similarity between the players' names. Hence for every player born on a certain date in one of the databases, we calculated the similarity between the names of every player born on the same date in the other database. To determine the best matches inside that subgroup, we used an iterative process of selection as seen in Figure 7.

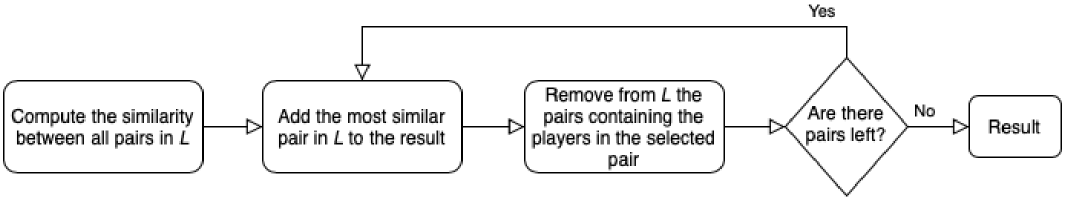

Yes
Compute the similarity between all pairs in *L*
Add the most similar pair in *L* to the result
Remove from *L* the pairs containing the players in the selected pair
Are there pairs left?
No
Result

Figure 7: Process followed to determine player matches between databases on a subgroup of players.

This method worked better than a specific optimization algorithm to solve an assignment problem such as the Hungarian Method, considering that most times the lists of names to compare contained different amounts of items. This difference is mostly caused by the fact that the FIFA database contains many more players that the shot subset of the OPTA database. Such a difference would have sometimes caused the optimization algorithm to add noise to the problem. After executing our algorithm, more than 90% of the shooters in our database were correctly paired to their equal in the FIFA dataset.

After a manual analysis of the remaining unpaired players, we discovered that some of them were not correctly paired because their dates of birth were unequal on both datasets. Most times the differences in the dates were of just one single digit, so we assumed those were errors produced when either OPTA or FIFA collected their data. To solve this, and bearing in mind that the number of remaining players to pair had reduced considerably, we executed again the process in Figure 7, but this time on all the remaining ones and not just those who shared their date of birth.

As mentioned before, making this iterative process on large lists of players induced some matching errors. So this final execution required a more detailed revision of the results, as well as a few manual corrections. The remaining unpaired players on the OPTA dataset mainly consisted on either very young players who typically played in U21 teams but had played a few matches in the first team or players from the third tier of Spanish football who played in the *Copa del Rey* competition. These players were not contained in the FIFA database, so it was not possible to access their ratings. We could have tried to interpolate them using maybe values from players who played in similar competitions and similar roles, but given that they were linked to very few shots in the dataset, we decided to just discard those shots.

The final variables we decided to keep from the FIFA dataset for each player are the following ones:

- **Overall**: Mean of all the player ratings.
- **Shooting**: Weighted mean of all the player ratings related to the shooting ability.
- **Finishing**: Rating from 0 to 100 of how accurate is the player shooting from inside the penalty area using the foot.
- **Long shots**: Rating from 0 to 100 of how accurate is the player shooting from outside the penalty area using the foot.
- **Heading**: Rating from 0 to 100 of how accurate is the player shooting using the head.
- **Weak foot**: Rating from 1 to 5 of how good is the player, in general, using his weak foot.

3.3 Dataset

The last step before starting to build the models will be to set up the dataset. We aim to group a set of informative features and as independent as possible. To summarise, the variables collected until now refer to the location of the shot, the context of the play and the player's abilities.

Instead of just using the location of the shot as XY coordinates, we will extract new features that will be much richer. The most basic one will be the Euclidean distance to the centre of the goal. A second one will be the angle between the line formed by the shot location and the centre of the goal and the goal line. We assume that the closer you are to the goal, the easier it is to score, but being in front of the goal will also increase your chances. Following this assumption, we will also add the angle between the lines formed by the shot location and each one of the goalposts. Finally, since we assume that it is equally likely to score from one lane or the other, we will convert the Y location variable so that it is 1 on the middle of the field and 0 on each sideline. We know that there are correlations between some of the produced and original positional features, but we will keep all of them because some families of classifiers benefit the additional information of non-linear relations.

The surrounding situation of a shot changes a lot depending on its preceding actions. For instance, a headshot coming from a corner will generally be more difficult to make given the higher density of people in the area than a header from a cross during open play. Tracking data would make this kind of information accessible, but having only event data it is much harder to account for it. Previous literature on this kind of models with event data [Kharrat et al., 2017] considered necessary using different classifiers for all these different situations. But that reduces the amount of available data for each classifier. That is why we will just focus on the shots produced during regular play — which is the most common type — and not consider the rest for the moment, at least until enough data is collected for each type to produce reliable classifiers. Regarding the rest of the regular play contextual categories, we will use one-hot encoding to store them.

The player ratings contain information about the players' abilities on specific kinds of shots. So we will keep the overall and shooting features as they are, but we will create a new one combining the other four. This new variable — we will call it *player ability* — will have the player's finishing rating on foot shots from inside the penalty area, the player's long shots rating for those from outside and the heading rating for headshots. If the foot shots are performed with the player's weak foot, the selected rating will be weighted by the player's weak foot rating — being a 5 the 100% of the original rating and a 1 only its 20%. We do this transformation because we suppose that the player's ratings on other kinds of shots are irrelevant for a specific kind of shot. Table 2 summarises all the variables included in the dataset.

Arguably, players recognize they should not shoot from very distant locations given that their chances of scoring will be practically 0. That is why we have very few shots on the first 2/3 of the field as well as close to the sidelines. But the models will need to get somehow this knowledge the players have to avoid assuming that no one has

Location-based	Contextual	Player-related
X position	From outside box	Foot shot
Y position centred	Without previous pass	Head shot
Goal distance	After rebound	Left foot shot
Goal angle	Second play	Right foot shot
Goal posts angle	Cross	Overall
	Through ball	Shooting
	Back pass	Player ability
	Pass from inside box	
	Pass from outside box	
	Drives to box	

Table 2: Variables used in the dataset divided into categories.

tried it before only because no one has ever thought of it. There are different ways to deal with this problem. The first one we tried was to remove from the dataset those goals scored from very distant locations, considering them to be outliers which were affecting the learning process of the models. But given that there was still a lot of missing data on those areas, some models still estimated high probabilities on shots from certain locations. We finally decided to randomly generate a few failed shots on those areas without data instead. The contextual variable for the shots was set to be the outside of the penalty area — since it is the one which would correspond — and we randomly produced the player ratings using the distribution of each variable. This approach helped to lower the probabilities produced by the models on distant regions, but we had to control the number of fake shots added so as not to increase significantly more the already imbalance of the dataset.

The last step will be to standardize the data to ensure that the models do not focus on certain variables just because of their range of values. We will do so by transforming each feature's values so that their mean is 0 and their standard deviation is 1. The final size of the dataset will be of 87161 shots, of which 8083 are goals.

3.4 Model Design

Before explaining the procedures followed to build the models, we will first need to formally describe the problem we are facing. We intend to learn a classifier able to model a Bernoulli random variable X — goal or not goal — given different types of information about the situation. This is best described by the following expression,

$$Expected\ Goals_t = \mathbb{E}[X|S_t, C_t, P, A_t = Shot]$$

where S_t corresponds to the location of the player in possession of the ball at time t, C_t to the contextual information also at time t, P to the information regarding the player's quality and A_t the chosen action to be performed by the player at time t. Given the Bernoulli nature of the variable X, the expected value expressed in the formula, when properly calibrated, may be interpreted as the probability of X being equal to one — or scoring a goal. It should be noted that this probability will only correspond if the player's action is a shot, as specified in the formula.

The fact that the expected value can be interpreted as a probability will be very useful because it will allow us to just sum the expected values of different shots to obtain the number of goals a team or a player was expected to score during a match or a whole season. And we will not be limited to know just the probabilities of shots becoming goals, but also any on-ball action provided that the player should have decided to shoot. For training the classifier though, we will only use shots since they will be the only actions to have explicitly tagged whether a goal was scored after a shot.

3.5 Model Implementation

From the previous definition, we can conclude that we need a binary classifier. Each entry on our shot dataset will contain information about the shot as well as a variable indicating whether it was a goal or not. That will be the target variable of the classifier. Furthermore, we need our classifier to yield a probability of belonging to the "goal" class.

It is difficult to know beforehand which kind of classifier will produce the best results for our particular problem, so we will choose three, each from a different family of classifiers, and pick the best one. The first choice will be the Logistic Regression, given that it is one of the most commonly used and it works well with linear dependencies. The second one will be XGBoost, a tree boosting system first presented in [Chen and Guestrin, 2016]. This kind of model scales very well to large datasets and has given good results on related research [Kharrat et al., 2017]. The third classifier we are going to use is an artificial neural network. The configuration that yielded the best results after testing a few different ones consists of two dense hidden layers with ReLU as the activation function, and the output layer with a single neuron with a sigmoid as the activation function. The sigmoid function will enable the ANN to produce a value between 0 and 1 as the output, which we will interpret as a probability.

Apart from selecting the kind of classifier which works best, we also want to build three different models, each one using different shot information, to observe the differences in performance. This will allow us to have separate interpretations, and to see which kinds of information helps best to predict whether a shot will be a goal or not. The first model will just contain data on the location of the shot — S_t in the previous formula —, the second will include the context, as well as the location — S_t and C_t in the formula —, and the final model will comprise all the available information.

Before learning each classifier for each model, we will partition the corresponding dataset into a training and test set, to see how well the classifier behaves on unseen data. When performing this partition we will apply stratification, to make sure that both sets contain the same proportion of goals. That will be key given that our target variable is imbalanced towards non-goals, and a non-stratified sampling could bias the classifier. We will also employ stratified partitions on the 10-fold cross-validation which we will perform for each combination of the classifier's hyper-parameters to tune them.

We will use the logarithmic loss (log loss) metric to evaluate the trained classifiers. We have chosen this metric because the output of our classifiers will be probabilities, and the log loss takes into account the confidence of the prediction when penalizing wrong predictions.

3.6 Model Results

After training all the classifiers mentioned in the previous section, the results in Table 3 were obtained. As stated, three different classifiers were tried on three different sets of variables of the same dataset. The evaluation metric used and shown on the table is the log loss.

The identifiers of the model column refer to the formula presented in the Model Design section (3.4), depicting which variables were used to build the model. Regarding the rest of the columns, we want to check for each classifier type the training

	Logistic Regression		XGBoost		Neural Network	
Model	Train	Test	Train	Test	Train	Test
S_t	0.2540	0.2608	0.2459	0.2577	0.2525	0.2601
S_t, C_t	0.2489	0.2564	0.2407	0.2538	0.2478	0.2571
S_t, C_t, P	0.2485	0.2558	0.2392	0.2536	0.2448	0.2554

Table 3: Summary of the results obtained for all the models and classifiers.

error they yielded for each set of variables and the test error. The training error will be useful to check whether the classifier overfitted or not. A significant difference between the training and test error would indicate so. The test error will be used to determine which is the best classifier for each set of variables of the dataset.

After taking a look at the table we can see that the training error values are always lower than their corresponding test error, but they are similar enough to assert that the model is not overfitting. The classifier with the lowest test errors is the XGBoost for all the variable sets.

All the classifiers tend to improve their classification power when more variables are added to the dataset. In all cases, the biggest improvement comes after adding the contextual information of the shots to the already existing location. Both the Logistic Regression and the Neural Network still produce a good jump in performance when adding the player information, whereas the XGBoost just improves a little. But given that it was already performing very well with the previous variables, it still yields the best results.

Although the improvement in performance is not as big as when adding the contextual information, there is an improvement, and having individual player information as model features will have other benefits. For instance, we will be able to compare the expected performance of several players given their shot qualities.

3.6.1 Calibration

The output of the classifiers we just trained are values between 0 and 1 and that is why we can interpret them as probabilities. But first, we have to make sure that they correspond to reliable probabilities. In other words, if the predicted xG of a shot is 0.1, we expect that 1 out of 10 times that shot will become a goal. But since each shot is only taken once, we won't be able to check that directly. Another way of doing this verification would be to group shots with similar xG values into bins, and control that their mean xG is similar to the percentage of goals within that bin. That is the idea behind the calibration plot presented in Figure 8. In this case, we chose 10 bins, which is a commonly used number.

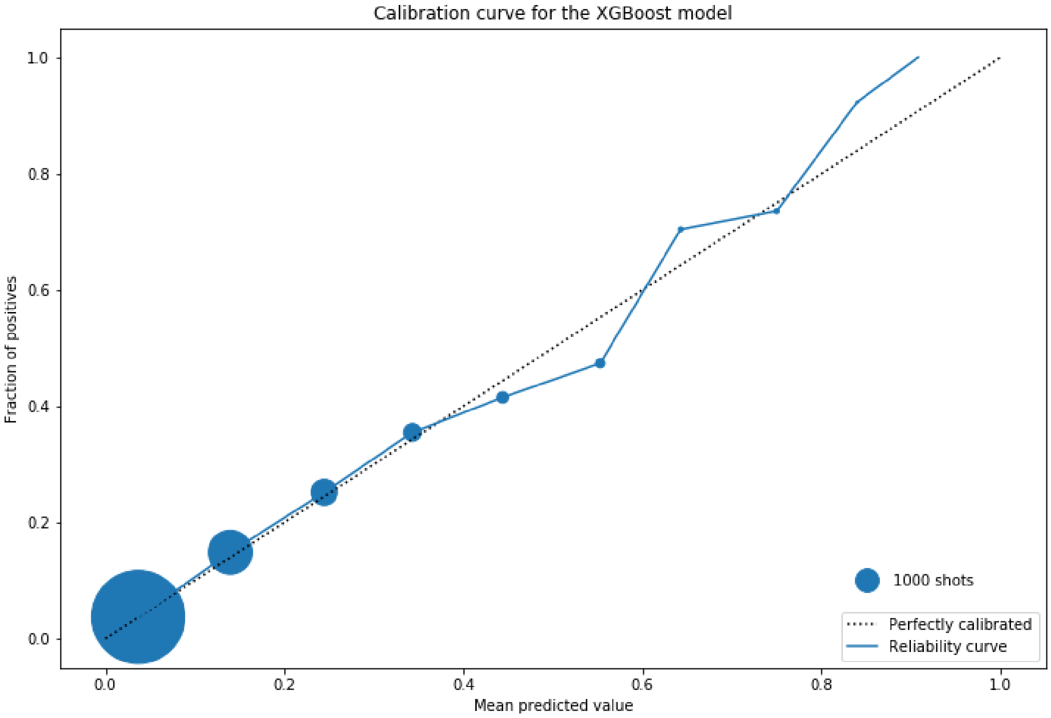


Figure 8: Calibration curve of the XGBoost model with all the features.

Provided that the X-axis corresponds to the mean predicted value and the Y-axis to the fraction of goals, a perfectly calibrated model would follow the dashed line where the values on both axes are always equal. In the case of the output of our best classifier, the curve generally follows closely the dashed line, until more or less the 0.5 predicted value where it starts to deviate a little. That also coincides with a significant diminishing in the sample size of the bins, indicated by the magnitude of the dots. The deviation is caused by the fact that we are using very few values to compute the mean value of those bins. The more values the bin has, the more will its mean approximate to the real calibration value. So it is a limitation of the plot. But we observe that despite this slight diversion, the tendency of the curve seems to follow the dashed line.

To see if an improvement can be achieved on the calibration of probabilities, we will try two of the most used post-processing calibration methods. The first one is called Isotonic regression and was first presented in [Zadrozny and Elkan, 2002]. It is a non-parametric method which learns a constant function by minimizing the square loss between its output and the actual target. The second one is the Platt scaling, introduced by [Platt et al., 1999]. This is a parametric approach given that it uses a logistic regression model to predict the target probabilities given the non-probabilistic predictions of the classifier to calibrate. Both post-processing methods require a validation set, which should be disjoint from the data used for fitting the classifier. We will use the same 10 folds previously used for the cross-validation, but this time the training set will be used to fit the classifier and the validation set to fit the calibration regressor. Given that we will now have 10 different couples of classifier and regressor, we will use the average of the predicted probabilities of the 10 couples as the final predicted probability.

We use the Expected Calibration Error (ECE) in Table 4 to present the results given by each of the tested calibration methods. This metric was presented in [Guo et al., 2017] and measures the difference in expectation between confidence and accuracy — they call it calibration gap. As in the calibration curve, it uses bins with equal ranges of predicted values and adds the calibration gap of each bin, weighted by the number of samples in the bin.

	ECE
No post-calibration	0.00417
Isotonic regression	0.00446
Platt scaling	0.01592

Table 4: ECE calibration metric value for each calibration method.

Considering that perfect calibration would correspond to an ECE value of 0, we observe that all the results obtained are very good. But it looks like none of the calibration methods tested improves the calibration of the probabilities. We assume that it is so because the classifier was already good in this respect and thus, it is hard to get a significant improvement out of post-calibration. The same tests were performed on the XGBoost classifiers fitted to fewer variables of the dataset — the ones presented in Table 3 — and similar results were obtained. Therefore we will stick to using the obtained classifiers without any post-processing steps.

3.7 Model Interpretation

As important as producing good predictions with reliable probabilities is, it will be hard to communicate them effectively unless they are interpretable. In the end, the results of our models will be used by players, coaches and analysts as one more source of information, and their usability will be limited if they are just a probability coming from a *black-box*. A very clear example would be when using the model on shots from players on their formative years. Knowing which features made their shots have lower probabilities of scoring will help them learn in which situations it is preferable to take a shot and in which it is not.

Generally, requiring interpretability limits the range of models you can use. Complex classifiers are often more accurate for most problems, but they tend to act as *black-boxes* and their results are more difficult to interpret. And that usually brings to a trade-off between accuracy and interpretability. We use SHAP, presented by [Lundberg and Lee, 2012], which assigns each feature an importance value for a particular prediction. An advantage of this approach is that it is a unified framework which works on any kind of predictor. It works by building a model called *explanation model* which attributes credit to each feature using several optimized methods.

In this section, we will focus on the importance the model gives to each feature used, as the SHAP framework also allows to measure it. An analysis of particular predictions using the framework will also be performed in Chapter 4. Both Figure 9 and Figure 10 summarise the impact each feature had on the output of the XGBoost with all the variables in the dataset. Figure 9 lists the features sorted by their importance. A distribution is shown for each feature, elongated on the X-axis of the plot. Each dot of the distribution represents a group of similar individuals — shots — analysed by the framework. The position in the X-axis quantifies the impact of those individuals on the model output — distinguishing between positive and negative. The colour indicates the value of the feature for those individuals. Figure 10 is a simplified version of the previous one, but it is more helpful to visually compare differences in average impact among the different features.

Both figures clearly state that the feature with the highest impact on the output by far is the goal distance. This makes a lot of sense provided that the more distant you are from the goal, the harder it is to aim there and the goalkeeper has

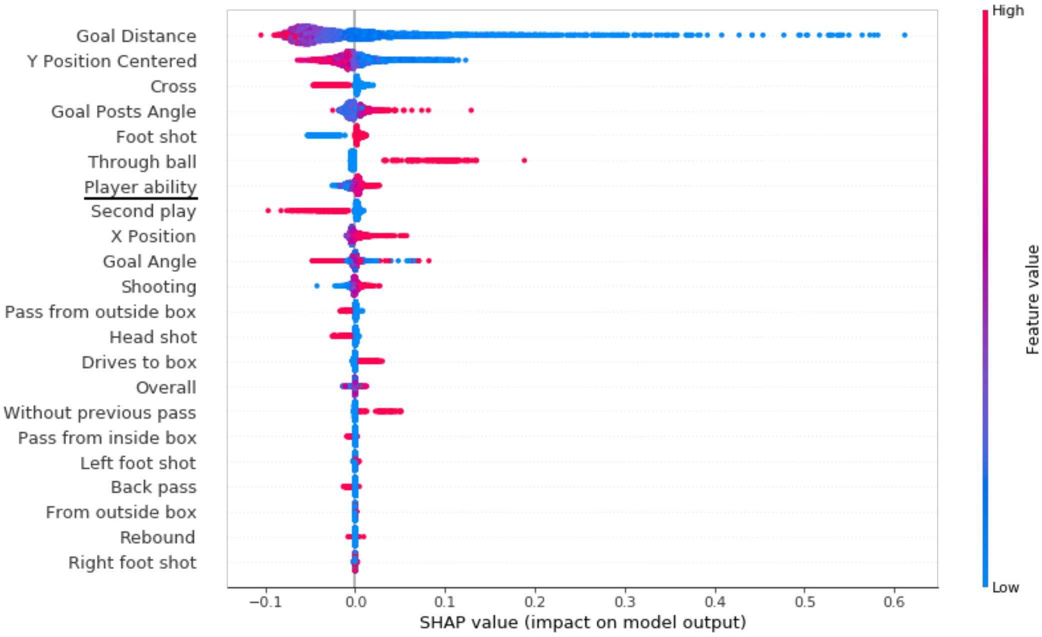


Figure 9: Feature importance on the final XGBoost model. Features are sorted from more important to less important.

more time to react to the shot. Plus, the ball will lose speed if it has to travel a long way. In fact, in Figure 9 we observe that when the goal distance has a very low value, it has a great positive impact, whereas high values have a moderate negative impact. A similar thing happens with the centred Y location, which is the second most important. This variable had very low values when the origin of the shot was centred in front of the goal, and higher values when it was closer to the sidelines. The model considers that being centred in the field will yield better chances of scoring than being to either one of the sides. The goal posts angle feature comes to a similar conclusion. The closer and more centred you are to the goal, the higher this angle will be, and the higher will be the positive angle on the outcome.

Then comes a group of features with a similar average impact which includes contextual variables as well as some regarding player quality. The model considers that the two most important contextual features are cross and through ball, the first one having a negative impact and the second one a positive one. Crosses are very common in football, but it is difficult to get a comfortable shot out of them. They require good coordination between the passer and the shooter, as well as a good ability from the passer to put it in the range of the shooter but inaccessible for

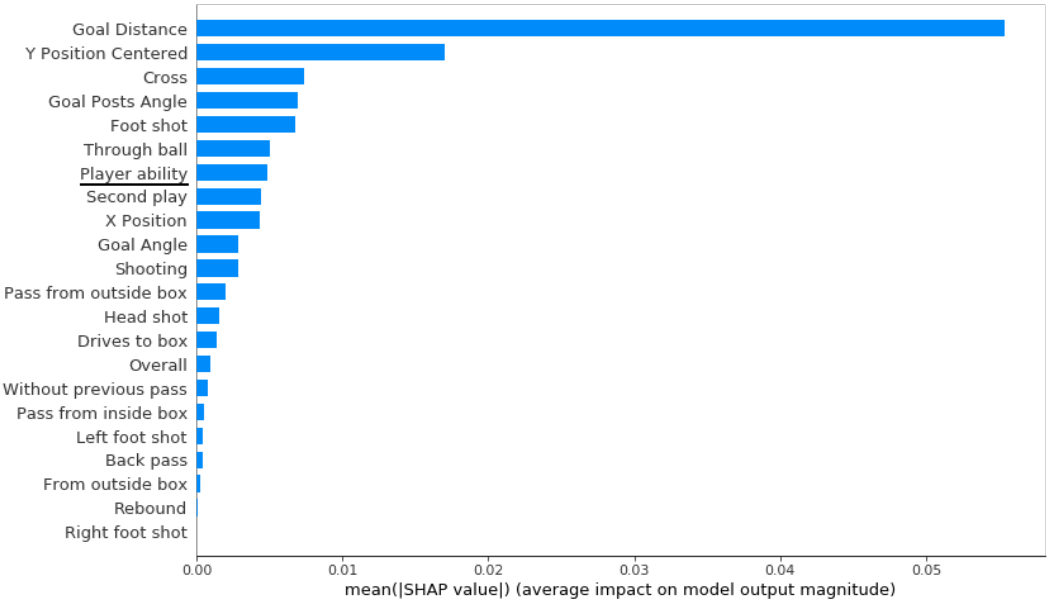


Figure 10: Simplified version of the feature importance on the final XGBoost model. Features are sorted from more important to less important.

the defenders. Besides, most crosses lead to a headshot, which is harder to make than foot shots according to the impact of those features in our model. On the other hand, through balls are passes that overcome the last line of defenders and normally lead to a one on one against the goalkeeper, which is a very favourable situation.

The player quality feature with the highest impact on the model output is the player ability. This variable was a mix of different player shot ratings which varied depending on some characteristics of the shot. As expected, higher values of player ability lead to a positive impact on the output whereas lower ratings lead to a negative impact. This feature falls on the 7th place of the ranking of feature importance, which indicates that being a talented player will help you score more goals, but other positional and contextual factors are more determinant.

The rest of the features have a lower impact on the model. Some of them even seem to have almost 0 average impact on the output. But removing them from the dataset slightly worsened the results of the classifier, so we decided to keep all of them.

4 Experiments

In this chapter, we will explore what can be done with the expected goals metric using the outputs of the 3 XGBoost classifiers we fitted. Since we have validated the models and made sure that the probabilities they yield are calibrated, we will assume that the predictions will be valid for both already seen and unseen data by the classifiers. Just as a reminder, the models we trained focused only on shots in regular play, and therefore the experiments will be limited to this kind of shots as well.

We will start by visually seeing how the location of the shots affects the probability predicted by the final XGBoost model. To do it, we will divide the pitch in small bins and compute the mean xG value of the shots that were produced inside the bin. We will then apply a KDE to smooth the areas instead of having just small squares. The result is presented in Figure 11.

We observe that the highest values are concentrated right in front of the goal, very close to it. That matches up with the conclusion we drew when checking the feature impact on the model, which was that the distance to goal was very important, especially with short distances. Then, rings are formed of diminishing probabilities the more you get away from the goal. We can see that the outer ring also has a very circular shape, which sustains the importance of the distance to the goal. But if we take a closer look, we can see that the highest values are condensed in very central locations, which also backs the importance given by the model to the Y location.

Note that some bins near the wings have slightly higher probability values than their surroundings. As explained previously, that is because it is very uncommon to see shots from there, and if by chance some of them were scored it can make the classifier think that it is a good location to shoot from. We solved that problem by adding failed shots from the wings to help the classifier learn that players don't usually shoot from there because it is very hard to score a goal. But adding too many shots would imbalance, even more, our dataset. The quantity we finally added was

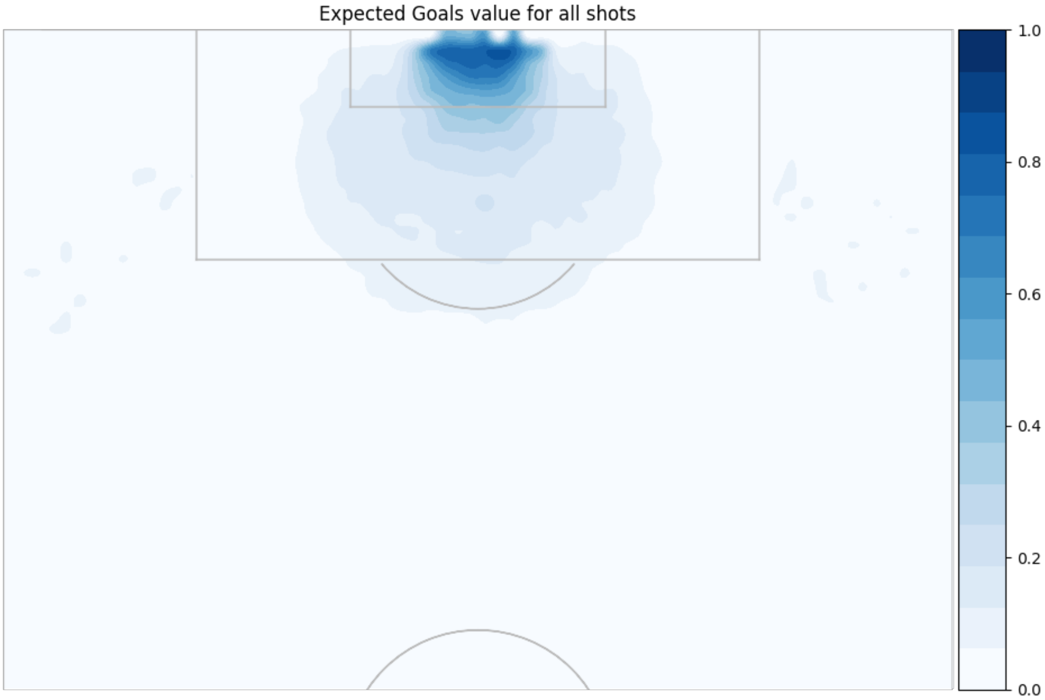


Figure 11: Distribution of the xG value along the field using the location of the shots in our dataset.

enough to keep the probabilities of these shots low, and although they are still a little bit higher than their surroundings, they are still values very close to 0.

If we now compare this plot to the one shown in Figure 12, we see a lot of similarities. That one illustrates where the goals were scored from, accumulated as a histogram where darker blue indicates the areas with a higher concentration of goals. Both maps have a similar shape, but the second one shows its peak slightly farther from the goal since it is less common to take shots from that close. But Figure 11 indicates that the few shots taken from there end up almost always in goal. The similarity between both figures reiterates the importance of the location of the shot when evaluating its outcome.

Figure 13 takes this analysis one step further. Both plots shown are similar to the one illustrated in Figure 11, but this time the player rating of the shots was modified before predicting the probabilities with the final model. In plot (a), instead of using the real ratings from the actual players who produced the shots in the dataset, we

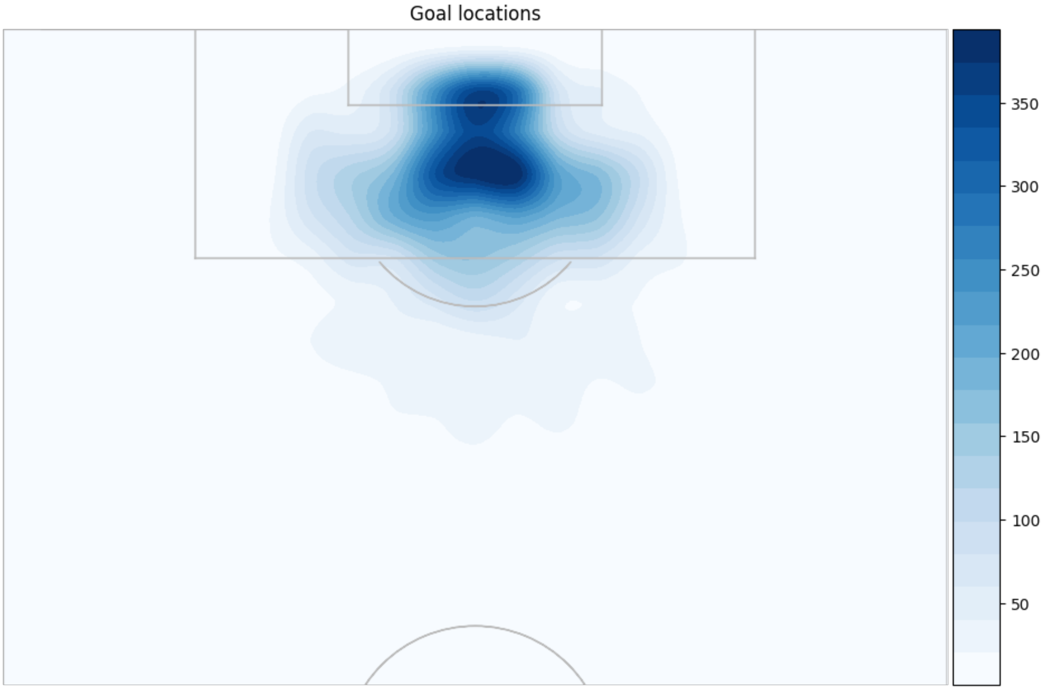
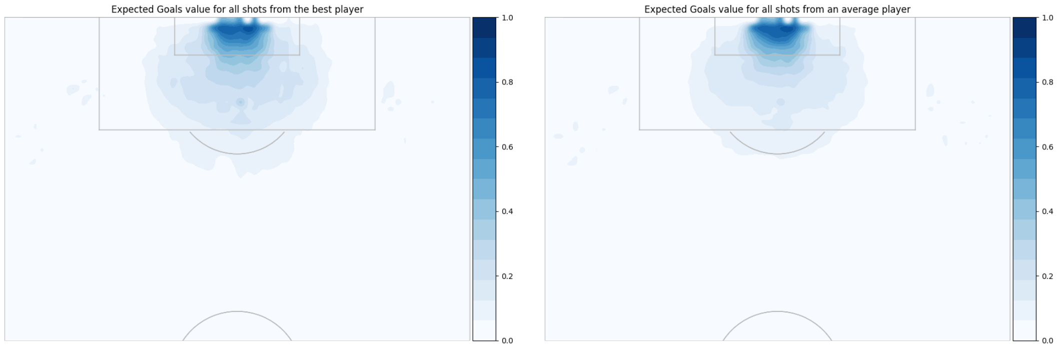


Figure 12: Histogram of the goals in the dataset by their location.

substituted them for the rating of the best player in each rating category. In other words, the values correspond to how the best player in the dataset is expected to perform on all shots. Plot (b) illustrates the same idea but with the average player in the dataset, given that this time we used the average rating for each category.

We observe that the shape of both maps resembles each other. That shows that the importance of the shot location is very present regardless of the quality of the player. But we also observe some differences in the distribution of the xG values. The outer ring appears to be slightly bigger for the best player, which indicates that he is expected to perform slightly better from long-range shots. This conclusion is also applicable to the rest of the outer rings, but the effect diminishes the closer you get to the location of the goal. This suggests that the importance of the player quality decreases for shots very close to the goal. It makes sense as those shots are expected to be easier to score, and average players should be able to do so.



(a) Player with the best stats

(b) Player with average stats

Figure 13: Comparison between the xG distribution of shots performed by the best rated player (a) versus an average player (b).

4.1 Single shot analysis

The next experiment we want to focus on corresponds to the continuation of the feature impact analysis we started in section 3.7. This time, instead of interpreting the outputs of the model as a whole, we will focus on single outputs. That will be crucial for players and coaches to understand what could be improved from the chances they have had.

Figure 14 shows the impact of the features on the prediction of 5 different shots Lionel Messi has tried within the last few seasons. Each shot is drawn in the first plot and it is paired to its corresponding impact plot by the number of the shot and the location of the impact plot from top to bottom. The shots 1, 2 and 5, which are marked in red, ended up becoming goals.

Regarding the impact plots, the features in red represent the ones which contribute to increasing the resulting probability, whereas the features in blue are the ones that decrease it. The length of the bar corresponding to a variable represents the level of impact of that variable onto the outcome. The longer the bar, the higher the impact. Note that the values right next to the feature names do not seem to make much sense. That is because they are not the original values of the variables, but the standardized version of the values which was fed to the classifier. So we will not pay much attention to them.

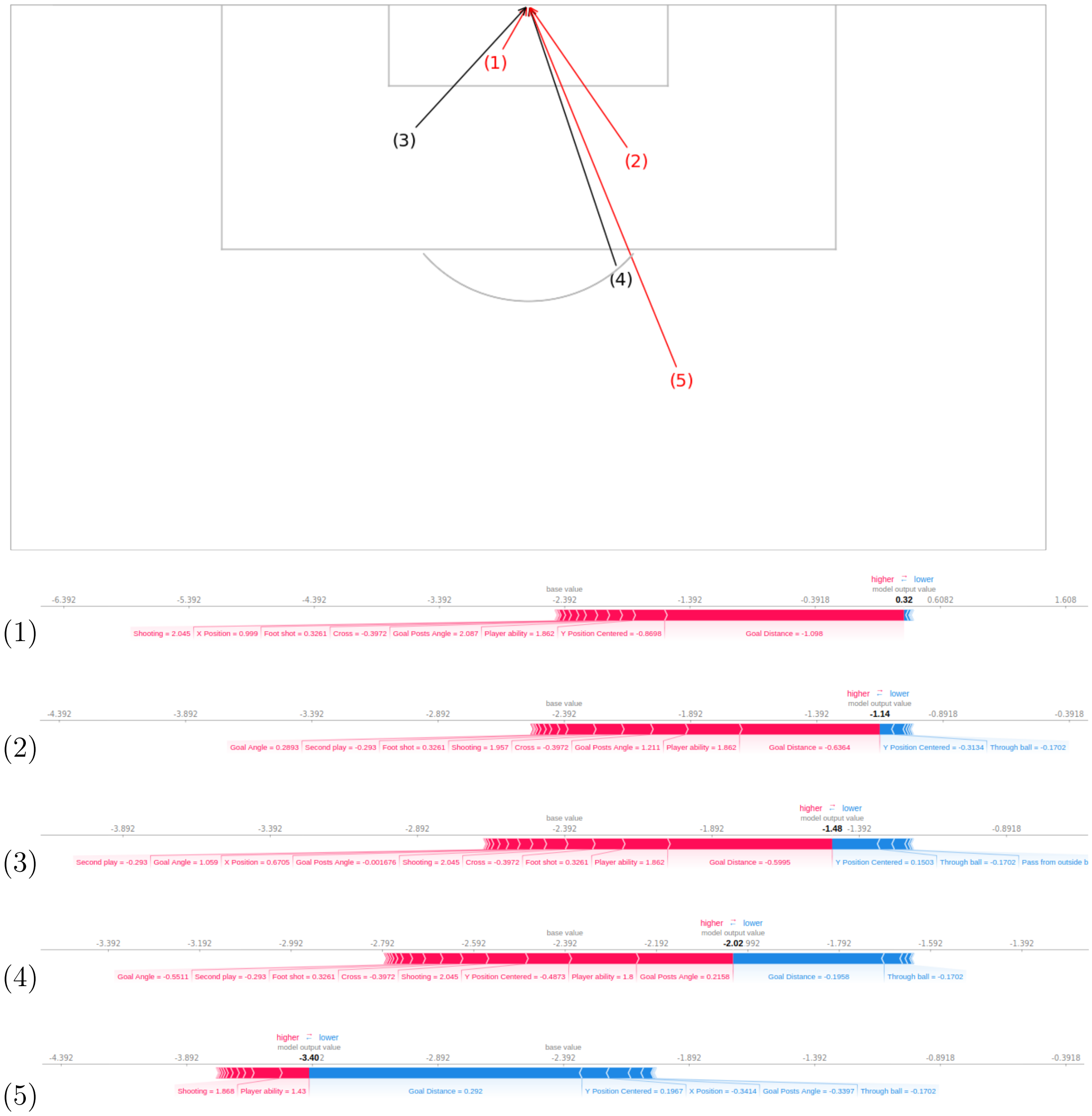


Figure 14: Feature impact on the xG of 5 different shots from Lionel Messi. The first plot illustrates the location were the shot was originated, and the plots below show the feature importance for each shot. The shot number corresponds to the location of the plots below from top to bottom. The red arrows correspond to shots ending in goals.

The first shot is the one with the highest resulting probability (0.58). It is evident given that most of the variables contribute to increasing the probability and barely any decreases it. This value could seem like a low probability, but it is not if we take into account the fact that the average xG value for all shots is about 0.1. That means that on average, you would need 10 shots to score a goal. Getting a probability almost 6 times higher than the average is not that common. We observe that the most important features in this particular shot are quite similar to the ones determined for the whole model. The fact that Messi has very good ratings and that the shot is very centred, both the Y location and the player ability features have the second and third highest impact.

The second and third shots are located in similar locations but on opposite sides of the field and have a predicted probability of 0.24 and 0.19 respectively. Although the second shot is slightly more distant to the goal, it is also somewhat more centred. This last observation is key since it is the Y location feature which makes the difference by significantly impacting more the probability on the third shot and making it decrease. Messi ended up scoring a goal with the second shot, but not with the third.

The last two shots are both from outside the penalty box, and so they have lower probabilities — 0.11 the fourth and 0.03 the fifth. This time, the feature which pushed the most the previous shot's probabilities towards being higher — the distance — has turned against them due to being far away. The fourth shot though, bearing in mind that it is quite centred in the goal location, still has several features impacting to increase the probability, such as the Y location and goal posts angle. It is not the case of the fifth shot, considering that it is more distant and less centred. The few variables that still impact towards increasing the probability are those regarding the player quality. In this case, Messi scored with the tougher one and missed the "easy" one.

4.2 Player performance analysis

As mentioned in section 3.4, the sum of the probabilities of a group of shots can be interpreted as the number of goals that are expected to be produced from those shots. So if we sum the probabilities of the shots performed by a player during a season, we will have the number of goals the player was expected to have scored. When compared to the actual number of goals, this value can be used as a metric for player performance.

Table 5 lists the ten top scorers from regular play during the first 27 fixtures of the 2019-20 season of the Spanish La Liga — just until the break caused by the pandemic of the COVID-19. For each player, we show their overall rating determined by the FIFA videogame, the number of shots in regular play, the sum of probabilities of those shots yielded by the models with and without the player information and finally the amount of them which became goals. The overall rating will serve as a measure of the quality of the player and the number of shots will indicate the number of opportunities a player has had to score goals.

Player	Overall	Shots	xG_{S_t,C_t}	$xG_{S_t,C_t,P}$	Goals
Lionel Messi	94	75	10.27	12.08	12
Gerard Moreno	81	59	8.55	8.86	11
Karim Benzema	87	84	11.22	11.65	10
Luis Suárez	89	46	6.08	7.02	9
Chimy Avila	75	62	6.32	6.23	9
Ante Budimir	72	32	5.53	5.53	9
Joselu	74	43	7.36	7.19	8
Loren Morón	80	34	5.78	5.87	8
Maxi Gómez	80	26	3.44	3.59	8
Ángel Rodríguez	79	33	3.29	3.27	8

Table 5: Open play shot stats for the top ten goalscorers of the first 27 fixtures of the 2019-20 season of the Spanish La Liga.

The expected goals values from both models can be interpreted as follows. The output of the model that does not use the player ratings is the expected number of goals an average player would score given the same opportunities. On the other hand, the values yielded by the model which includes the player ratings correspond to the expected number of goals a player with a similar quality would score with the same opportunities.

So when analysing the Table 5, we can point out that the players that have scored more goals than they were expected are over-performing, whereas those with fewer goals than expected goals are under-performing. Among the players shown, only one — Karim Benzema — seems to be under-performing. That is because it is easier to be part of the top ten scorers if you are an over-performer than if you are an under-performer. In the case of Lionel Messi, we see that he is over-performing in comparison to an average player, but he is doing exactly as expected for a player with his qualities. Karim Benzema is the third in the list, but he is also the one who took more shots and the one with the highest accumulated xG_{S_t,C_T}. Curiously enough, Messi is the one who was expected to score more goals accounting for the player quality.

When looking at the rest of the table, we see some players with relatively few shots and low accumulated xG. They are generally either player from teams with fewer offensive power and therefore, with less goal-scoring opportunities or players who have played fewer minutes due to injuries, or because they normally start from the bench. Despite these adversities, they managed to score a remarkable amount of goals. To detect these over-performing players we can use metrics such as the difference between the number of goals and the accumulated xG, or the average xG value per shot — accumulated xG divided by the number of shots. The first one will be a measure of performance, whereas the second will indicate whether the player tends to have very clear opportunities or difficult shots.

In the case of the $xG_{S_t,C_T,P}$, those players who have performed better than their estimated qualities in a regular manner during a season, should be expected to get a better rating for the following year. That will cause his expectations for the next year to grow accordingly, and he will be expected to perform at the level of the previous season. A similar adoption would also be expected for the under-performing players.

4.3 Team performance analysis

We performed a similar analysis with teams. Table 6 shows the accumulated xG of the teams of the Spanish La Liga for the same period as Table 5. They are sorted by the number of goals they managed to score in open play, and additional information is included such as the number of shots they took and the position on the table they found themselves at the end of that period.

The accumulated xG of a team can also be seen as the danger that the team produces from shots. That is because the accumulated value will account for both the frequency of shots — the more shots you take, the more values you will sum together — as well as the quality of the shots — the clearer the opportunities, the higher will be the values you will sum. So the teams which will be considered as offensively powerful will be those with a good combination of good and frequent opportunities.

Team	Position	Shots	xG_{S_t,C_t}	$xG_{S_t,C_t,P}$	Goals
Barcelona	1	262	35.92	39.51	47
Real Madrid	2	350	37.55	38.57	37
Real Sociedad	4	248	27.98	28.24	33
Villarreal	8	309	32.15	33.05	32
Real Betis	12	264	23.94	24.65	27
Sevilla	3	262	29.50	29.95	26
Atlético de Madrid	6	246	30.55	31.41	25
Levante	13	254	24.49	24.91	25
Valencia CF	7	182	20.90	21.50	25
Getafe	5	208	19.15	18.94	23
Osasuna	11	260	23.92	23.32	22
Granada CF	9	205	18.48	18.02	21
Alavés	14	153	18.29	18.05	19
Athletic Club	10	211	16.70	16.91	18
Mallorca	18	177	16.55	15.99	17
Real Valladolid	15	205	18.46	18.73	16
Eibar	16	206	16.82	16.98	16
Leganés	19	223	20.02	19.49	15
Celta de Vigo	17	203	20.13	20.88	14
Espanyol	20	232	21.58	21.03	12

Table 6: Team shot stats from the first 27 fixtures of the 2019-20 season of the Spanish La Liga. The teams are sorted by the number of goals they scored.

We observe that the number of goals scored and accumulated xG correlates with the position in the table, but there are clear exceptions. Take as an example the two teams from the city of Seville: Real Betis and Sevilla. Both teams have roughly the same number of goals and shots, but Sevilla has a much higher accumulated xG. Having the same number of shots but higher accumulated xG indicates that their shots have on average a higher chance of becoming goals than Real Betis'. But we also see that there is a big gap between both teams in their position in the table. Both Sevilla and Betis are offensively powerful teams, and we find them one right next to the other in the table because we are just focusing on the offensive side of the teams. The main difference between these two teams is that Sevilla is also a very solid team in defence, whereas Betis not so much.

The accumulated expected goals can also be used to evaluate defensive performance, though. If instead of adding the expected goals of the shots the team made we add the shots the team received, we will have a sense of the danger allowed by the team and performed by its rivals. If we checked those metrics we would see that Betis and Sevilla differ in that sense.

As it happened with Messi and Benzema, Real Madrid has a higher expectation of goals if they were performed by average players, whereas Barcelona wins when accounting for the quality of the players. And so we see that Barcelona performed much better than both the expectation of average players and players with their qualities, but Real Madrid performed as expected with average players, but given the qualities of their players they were supposed to do a little better.

Finally, in the lower part of the table, we see that the last teams in the classification are not the ones with the worst offensive production. But they are the ones with the worst difference between goals and expected goals. So we can use this metric with teams as well to measure performance.

5 Conclusions and future work

In this project we have presented two different models that allow us to predict the probability that a shot becomes a goal, leading to different interpretations of the probabilities. The first model was trained without using any specific information of the player taking the shot, so considering that the shots the model was trained with were taken a wide variety of players, we can assume that the probability corresponds to being shot by an average player. The second model did contain information about the shooting quality of the player, so the probability will correspond to being shot by a player with their qualities.

We have explained the importance of the calibration of these probabilities, and have also proven that they were indeed calibrated. We have also shown how the probabilities can be combined, and presented various metrics derived from them which help to measure both team and player performance.

We have also mentioned how relevant it was for the results and the models to be interpretable. We have shown a framework which enables us to know both the importance of the model features as well as the decisions made to yield the probabilities of sample shots.

During the development of the project, we have realised how important domain knowledge is. It probably applies to most domains, but given the complexity of football, it would have been very hard to determine which factors can affect producing goals if it were not for the help of the experts. Their aid has also been very useful when facing inconsistencies in the data. But they have also benefited from their collaboration with this project. They claim that the process of designing the algorithms to create the different football concepts has helped them to build a more objective conception of the knowledge they already had, instead of relying so much on their instincts. The results and interpretability of the models had a similar effect on them, as it allowed them to objectively quantify the importance of the different factors that surround scoring goals through shots.

5.1 Future work

The limitations of this project at the moment of the development — both temporal and data-wise — can be improved straightforwardly. First of all, if more seasons of the same kind of event data become available, they could be used to retrain the models to see if more data improve their predictive power. More types of classifiers and more configurations for the ANN that were not considered in this project due to time limitations can also be tested. If apart from more event data, tracking data also becomes available, it would be nice to see how much do the models improve with positional-based features, as well as how important they are for the models. Finally, with more data, we could also produce a custom system to attribute quality ratings to players.

6 Acknowledgments

First and foremost, I would like to thank both the tutor and the advisor of this project. They have been a great help in suggesting new ideas and modifications. I would also like to thank FC Barcelona and in particular the Sports Analytics department for their collaboration with this project. And finally, I would like to give special thanks to all the football analysts and coaches from FC Barcelona who have contributed to the project by sharing their knowledge.

References

[Blog, 2020] Blog, M. B. (2020). Mlblogs best blog. `https://mlblogsbestblog.wordpress.com`.

[Bransen and Van Haaren, 2020] Bransen, L. and Van Haaren, J. (2020). Player Chemistry: Striving for a Perfectly Balanced Soccer Team. *MIT Sloan Sports Analytics Conference*, pages 1–24.

[Cervone et al., 2014] Cervone, D., D'Amour, A., Bornn, L., and Goldsberry, K. (2014). POINTWISE: Predicting Points and Valuing Decisions in Real Time with NBA Optical Tracking Data. *SLOAN Sports Analytics Conference*, pages 1–9.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu:785–794.

[Decroos et al., 2019] Decroos, T., Van Haaren, J., Bransen, L., and Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (1):1851–1861.

[Decroos et al., 2017] Decroos, T., Van Haaren, J., Dzyuba, V., and Davis, J. (2017). STARSS: A spatio-temporal action rating system for soccer. *CEUR Workshop Proceedings*, 1971:11–20.

[Eggels et al., 2016] Eggels, H., Van Elk, R., and Pechenizkiy, M. (2016). Explaining soccer match outcomes with goal scoring opportunities predictive analytics. *CEUR Workshop Proceedings*, 1842:1–10.

[Fernández et al., 2019] Fernández, J., Bornn, L., and Cervone, D. (2019). Decomposing the Immeasurable Sport: A deep learning expected possession value framework for soccer. *MIT Sloan Sports Analytics Conference*, pages 1–18.

[Goldsberry, 2019] Goldsberry, K. (2019). *SprawlBall: A Visual Tour of the New Era of the NBA*. HOUGHTON MIFFLIN.

[Guo et al., 2017] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. *34th International Conference on Machine Learning, ICML 2017*, 3:2130–2143.

[Gyarmati and Stanojevic, 2016] Gyarmati, L. and Stanojevic, R. (2016). QPass: a Merit-based Evaluation of Soccer Passes.

[Kharrat et al., 2017] Kharrat, T., Peña, J. L., and McHale, I. (2017). Plus-Minus Player Ratings for Soccer. pages 1–17.

[Kram, 2019] Kram, Z. (2019). The 3-point boom is far from over. `https://www.theringer.com/nba/2019/2/27/18240583/3-point-boom-nba-daryl-morey`.

[Leone, 2019] Leone, S. (2019). Fifa 20 complete player dataset. `https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset`.

[Levenshtein, 1966] Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

[Lewis, 2004] Lewis, M. (2004). *Moneyball: The Art of Winning an Unfair Game*. W. W. Norton.

[Link et al., 2016] Link, D., Lang, S., and Seidenschwarz, P. (2016). Real time quantification of dangerousity in football using spatiotemporal tracking data. *PLoS ONE*, 11(12):1–16.

[Llana et al., 2020] Llana, S., Madrero, P., and Fernández, J. (2020). The right place at the right time: Advanced off-ball metrics for exploiting an opponent's spatial weaknesses in soccer. *MIT Sloan Sports Analytics Conference*, pages 1–16.

[Lucey et al., 2014] Lucey, P., Bialkowski, A., Monfort, M., Carr, P., and Matthews, I. (2014). "Quality vs Quantity": Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data. *Proc. 8th Annual MIT Sloan Sports Analytics Conference*, pages 1–9.

[Lundberg and Lee, 2012] Lundberg, S. M. and Lee, S.-I. (2012). A Unified Approach to Interpreting Model Predictions Scott. *Nips*, 16(3):426–430.

[Murphy, 2019] Murphy, R. (2019). Fifa player ratings explained: How are the card number stats decided? `https://www.goal.com/en-ae/`

news/fifa-player-ratings-explained-how-are-the-card-number-stats/1hszd2fgr7wgf1n2b2yjdpgynu.

[Platt et al., 1999] Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.

[Power et al., 2017] Power, P., Ruiz, H., Wei, X., and Lucey, P. (2017). "Not all passes are created equal:" Objectively measuring the risk and reward of passes in soccer from tracking data. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Part F1296:1605–1613.

[Ratcliff and Metzener, 1988] Ratcliff, J. W. and Metzener, D. E. (1988). Pattern-matching-the gestalt approach. *Dr Dobbs Journal*, 13(7):46.

[Rathke, 2017] Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise*, 12(Proc2).

[Sill, 2010] Sill, J. (2010). Improved NBA Adjusted + / - Using Regularization and Out-of-Sample Testing. *MIT Sloan Sports Analytics Conference*, pages 1–7.

[Spearman, 2018] Spearman, W. (2018). Beyond Expected Goals. *12th Annual MIT Sloan Sports Analytics Conference*, (August):1–17.

[Sports,] Sports, E. Fifa players. https://sofifa.com/.

[Sports, 2020a] Sports, M. (2020a). Metrica sports. https://metrica-sports.com.

[Sports, 2020b] Sports, M. (2020b). Metrica sports sample data. https://github.com/metrica-sports/sample-data.

[Statsbomb, 2020] Statsbomb (2020). Statsbomb open data. https://github.com/statsbomb/open-data.

[Zadrozny and Elkan, 2002] Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699.