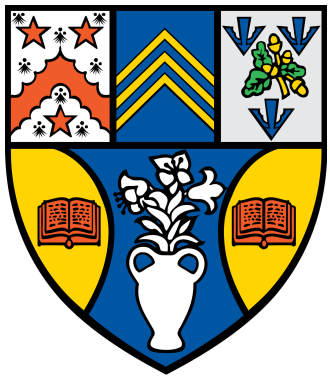
Combining Game-Engine Technology, Game-Based Learning and Sports Data Analytics to Reinforce Tactical Decisions in Football

Lewis Trainer

MSc Computer Games Technology, 2022



School of Design and Informatics
University of Abertay Dundee

Contents

List of Figures

List of Tables

Abstract

In modern football, there are very few successful top level football teams that do not utilise data analysis. Data is analysed by trained analysts and coaches, who act as a third party to relay information to coaches and players. There is a barrier between this data and its utility and the players and coaches who do not come from a scientific background.

The aim of this project was to develop an application that combines open-source footballing data from real matches and game-engine technology to instruct the player in situation specific football tactics. Data will be modelled and used as the main mechanism for formulating a score system used to assess player decision making within the game. A review system will be developed within the application to reinforce these decisions through the data analysis and display of the real-world footballing data used to assess the application user's performance as they progress.

The results of this research provide an initial framework for the combination of game-based technology and football data analytics, specifically for the scenario of choosing passing options within a game. The work conducted in this project and the subsequent outcome prove that combining these fields is feasible. With the addition of game-based learning tools this work has the potential to reinforce the tactics underlying the data models used to form mechanics within the game. However, in order to assess the strength of the tools developed within the application, as well as the practical application of the technology, further research is required.

1 Introduction

Data analysis and the use of analytical technology is growing in prominence in football (Herberger & Litke 2021). In particular, the use of situational video analysis and data collection is at the forefront. Whilst this technology is growing in popularity, there is still a lack of accessibility to this science for players and coaches who do not come from an analytical background, as much of the data requires technical knowledge in statistics and data science.

At the elite level, football players will only get a limited time(<1s) to make decisions, and often they will be under intense pressure, both physically and mentally, when doing so (of Sport 2022). One of the purposes of tactical training, in any sport, is to ingrain repetitive movements, patterns and decisions into a player's mind. This is so that when in high-pressure (both physiological and time based) situations, the player's subconscious takes over and effective decisions are made, minimising stress on the decision maker. In modern football, top level players are often exposed to excessive workloads, due to the increasing number of international breaks and mid-week games, resulting in less time on the training pitch, and more time in the 'recovery room' (FIFPro 2022). As a result, coaches and managers are limited in the amount of time they have to impose a specific tactic to a player or team.

This project aims to investigate whether it is possible to develop a technology that can combine football data analytics with game-engine technology and game-based learning tools, to enhance tactical understanding of the game whilst aiding maximisation of productivity of players and coaches, without hindering current practises and schedules.

Similarly to physical training routines, the repetitive nature of video games has been widely proven to enhance the performance of players the more they play the game (Gazzard & Peacock 2011). It is in fact extremely hard to find any game (at least one that does not use adaptive difficulty based artificial intelligence) in which the players performance declines as their hours of playtime increases. Some of the industry trailblazers such as Tetris (*Tetris* 1984) and Pac-man (*Pac-man* 1980) used simple tools such as point scoring in order to achieve this motivation and direction on how to maximise performance within the game, as you are judged on your ability to maximise scores per game. In the context of these games, and many like them, the purpose of score keeping and other similar tools is to help master the game. But what if the game was designed in a way that mastering it would enhance a tactical understanding of a footballing scenario, and act as a learning tool?

For some time, game-based learning has been used to assist learning (Faiella & Ricciardi 2015). It is a widely accepted phenomenon that many animals, including humans, have the inherent natural capacity to learn effectively through play (Bateson 2005), so it is no surprise that game-based learning has become a highly researched teaching method

(Breuer 2018). Gamification is already a utilised learning tool within the sport of football, with top clubs such a F.C. Barcelona utilising gamification tools throughout the 2020 lockdown, in order to keep players motivated whilst not having access to training facilities (The Barça Innovation Hub Team 2020). As a result, there is reason to believe that if gamification and learning tools such as interactive environments, visual contextualisation and tasks that are rewarded via point scoring are combined, then they could effectively be used to inform the user on a learning topic that may be otherwise hard to grasp (Hamari et al. 2014).

The aim of this project is to develop an application that utilises fast action, and infinitely repeatable simulated gamified scenarios, in an environment that does not require the player to increase their physical workload. Provided that the scenarios contained within the application are designed in a way that would enhance a players tactical understanding in a specific context, then there is reason to believe that such a technology could be used to supplement training done on the pitch, whilst players are getting the required physical rest needed to continue to perform at peak levels.

The application will use a public dataset of spatio-temporal match events in top level football competitions (Pappalardo et al. 2019*a*). This data will be modelled using methods and techniques that are currently used within modern football data analytics, and specifically models that are used to better understand tactical aspects of the sport. This modelled data will then be deployed in a game-engine environment that will reward the user for making optimal tactical decisions, based on the data that is displayed to them as they progress.

In order to provide genuine football education, this project attempts to combine game-engine simulation with footballing data from real matches, to show the effectiveness of various decisions made in games and relay this information to the user. This project should demonstrate a blank canvas of technology framework, that when combined with the tactical knowledge and understanding of an elite level coach and football data scientist, can be used in hyper specific scenarios to supplement training routines and match preparation. Data models will be developed based on the data available to the public, and will be designed with a game environment in mind. As the user of the developed application progresses, they will be able to assess their performance in game through a combination of review tools. A score system will be designed within the game-engine that combines the data models developed. The resulting scoring system implemented aims to motivate the player into maximising their performance based on the specific data models incorporated into the score. The review system will allow users to understand the reasons behind the score they received within the application, from both a gameplay perspective and through providing a statistical breakdown of the data that forms the score they receive for their actions.

The literature review (Sections 2 and 3) contained within this document will outline

the prevailing research within the game-based learning and gamification fields, providing context for the methodology used to apply these tools across disciplines into desired fields. The literature review also provides an overview of the impact of data analytics within sport in recent history. This section will also cover some of the common applications of data analytics within football specifically, particularly the tactical reasoning and uses of models used in the practical development of this project.

The methodology and application design section (Section 4) within this document will firstly outline the design process taken in the development of the final application, outlining the method taken to combine the various components of data modelling, game-based learning and game-engine technology. This section will then outline the practical processes used in developing the data models that form the scoring system used within the application, as well as the techniques used within the game-engine technology to create an interactive environment which combines all aspects of the design process, including the development of a review system which aims to allow a user to assess their performance, whilst understanding the data that is used to create the scoring system used to evaluate the users actions.

The results section (Section 5) outlines the results of the development of the proposed application, assessing the development in three stages. Firstly, the development of the data models used to form the scoring system within the application. Secondly the development of the main gameplay loop used within the application, as well as the incorporation of the data models in the game-engine. And finally the review system built to allow users to assess performance. All stages will be assessed on their practical usability within the purposes of the application, as well as the overall connectivity and interaction between each separate aspect of development.

Finally, the conclusion (Section 6) will summarise the work completed throughout this project and assess where the investigation was conclusive and if not outlines the potential future work that could be taken to deepen the understanding of how game-engine technology, game-based learning and football data-analytics can combine.

2 Game-Based Learning and Gamification

Throughout this project, game-based learning and gamification techniques will be combined to develop an application that informs users of potential tactics (i.e. optimal decisions in specific scenarios) and aims to reinforce this through incorporation of real data gathered across elite level competition.

The use of gamification tools and game-based learning are methods of implementing game-like features or gameplay elements to enhance learning. Both techniques share common attributes Figure 1, and in practise, the differences between the two is often hard to define. By definition the two techniques are similar: Game-based learning can

be defined as the borrowing of certain gaming principles, and applying them to real-life settings to engage users with defined learning outcomes (Pho & Dinscore 2015) (Plass et al. 2015). Gamification can be defined as "the use of game design elements in non-game contexts...to make non-game products, services, or applications, more enjoyable, motivating, and or engaging to use." (Deterding et al. 2011). Due to the similarity of formal definitions it is important to clarify the important differences between the two, and how that effects their applications.

An appropriate way to consider the differences between the two may be to think of game-based learning as a game designed with the desired outcome of teaching, whereas gamification is taking game elements to a non-game environment and using them as a tool in the learning process.

Figure 1: A diagram of the relationship between gamification, serious games and playful interactions (Deterding et al. 2011)

Hamari et al. define gamification as "a process of enhancing services with (motivational) affordances in order to invoke gameful experiences and further behavioural outcomes" (Hamari et al. 2014). A game design element is any design element used that is predominantly characteristic to, and often originates from, games (Deterding et al. 2011). Common examples of gamification elements (motivational affordances) would be a badge-based rewards system, or a point-scoring system that rewards success within the confines of the game rules and mechanics in question.

In recent years, gamification has seen a steep rise in interest from the commercial world, particularly within the marketing industry. Gamification has been observed to enhance user engagement and promote positive patterns when using services, such as

social interaction and user activity (Hamari et al. 2014). Due to the increased commercial success observed from the increased use of gamification, there has since been an increase in the volume of academic researched published around the subject, with Hamari et al. stating that the appearance of the term "gamification" itself has been increasing rapidly in titles of research papers, from circa. 2010 on-wards (Hamari et al. 2014). This demonstrates that gamification is growing in popularity as a main topic of research in the academic community.

Gamification and game-based learning, can cover a broad scope of disciplines. As such, in academic settings, it's effects have been studied across many disciplines and applications and have thus returned results lacking in empirical statistics on the overall effectiveness of gamification across disciplines. I.e. the application and measurement of the effectiveness of gamification/game-based learning techniques in a marketing context looks very different to an educational context. Whilst Hamari et al. attempted to perform an extensive literature review of empirical studies performed on Gamification (Hamari et al. 2014), this study was carried out in 2014 and since, the applications and research surrounding gamification has grown substantially. In order to pinpoint effective strategies of implementation of gamification it is critical to look in specific disciplines. As the application being developed in this project evolves around using a combination of gamification tools and game-based learning in its development, the effects of gamification in learning were researched. In 2020, Sailer and Homner attempted to build upon the works of Hamari et al. (Hamari et al. 2014) and Seaborn and Fels (Seaborn & Fels 2015) and produce a conclusive meta analysis on the effectiveness of gamification, specifically in the context of learning and education, focusing on the cognitive, motivational and behavioural learning outcomes (Sailer & Homner 2020). The cognitive outcomes of successful gamification are yet to be resolved. We can assume that a serious athlete would already have the external motivation required to use the application to learn. As a result we will be focusing on the desired behavioural outcomes of our gamified scenario, which in this case is related to improving the tactical awareness of the player based on the data portrayed to them. Sailer and Homner (Sailer & Homner 2020) found that the inclusion of game fiction and competition combined with collaboration, were particularly effective game elements for influencing behavioral outcomes of users. The gamification aspect of this project is centred specifically around the learning involved in understanding the data science that will provide the scoring mechanics within our application. Using gamification tools when displaying data to the user in a game-based learning environment should increase the players potential to understand why certain decisions are rewarded more greatly than others. Players will be able to review choices made throughout the application loop, and view not only the scores granted to them by decisions made, but also the scores available through other decisions, with a break down of how these scores were decided from a data point of view. As for the main design of the application, the football

scenarios contained within the game are designed to somewhat replicate scenarios that a player may find themselves in a competitive football match, thus providing an already game-based environment in which to employ our various learning tools.

A notable tool that could potentially be used to assess player performance would be to perform an Action-Cost analysis developed by Kavanagh and Miller (Kavanagh & Miller 2021). Kavanagh and Miller state that by considering the costs of actions (the value by which the player reduces their maximum probability of winning) of a player, then we can obtain an accurate representation of how skilful they are. In the context of this project, the action cost of the player would be consistent with the degree to which their decisions impacted their maximal probability of success (or the proximity to the optimal decision) and the skilfulness of the player would be assessed as to whether they are minimising the cost of their actions by the greatest amount possible (as the higher the score given to the player per action, the better the decision). An effect of the action-cost analysis is it would allow for the tracking of how players are learning over time, and also allow for analysis of which scenarios players are finding difficult to better understand why this would be the case. As the cost-action analysis allows the examination of how player skill changes over time, this would be a reasonable analysis to perform on a test group to determine the effectiveness of the game-based learning tools afforded to them throughout the application. It is worth noting that in the application developed for this project, a cost-action analysis would not necessarily have any implications on whether or not the user is gaining a better understanding of real match scenarios as the data used to construct scores is somewhat arbitrary, and is simply there to provide a reasonable context to decisions being made by the user.

3 Data Science and Analytics in Sport

Data science and analytics has been rapidly growing in sport over the last two decades. The first known case of a sporting team using analytics as a mainstream technique in decision making was the American baseball team, Oakland Athletics, whose story were told in the book "Moneyball: The Art of Winning an Unfair Game" by Michael Lewis (the same story portrayed and made famous by the popular film "Moneyball") (Lewis 2004). In the early 2000s, Oakland Athletics baseball team were the first known sporting team to take analytically proven statistics (that most other teams had deemed unimportant) to gain a competitive advantage over their opponents in player recruiting and in-game decision making, whilst doing so at a greatly reduced financial cost to their rivals. Due to the nature of Oakland Athletics's success, almost all rivals within the league and the sport soon began to invest and restructure their organisations to incorporate similar levels of data analytics and the techniques involved.

Since then, data analytics gradually spread throughout elite level sport, with the next

great impact being made in the NBA league in basketball, where the sport has undergone a process that has been coined as *the three-point revolution*. As almost every team in the NBA has a substantial data analytics department (Goldsberry 2019), it was discovered that it was not a favourable decision for teams to make mid-range shots if a three point shot option was available. Figure 2 shows that the average shot success percentage only changed as little as 5% between 3 and 25 feet. If we then take into account that beyond the 3-point line (between 22 and 23 feet) the points gained for scoring is increased by one, then we can notice that the average points per shot attempt rises above the expected values, as can be seen in Figure 3.

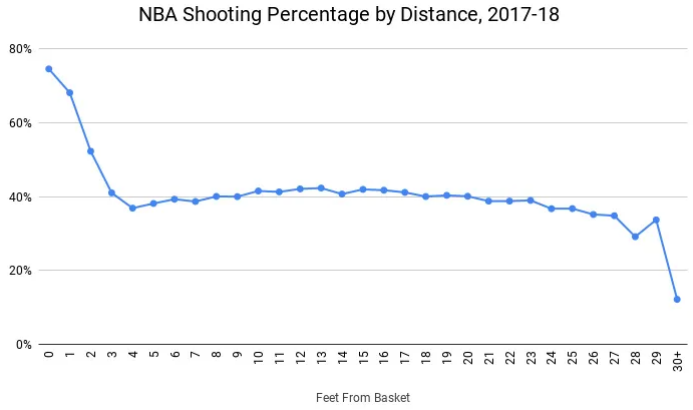


Figure 2: NBA shooting percentage by distance, using tracking data from NBA Savant for 2017-18 season. (Kram 2019)

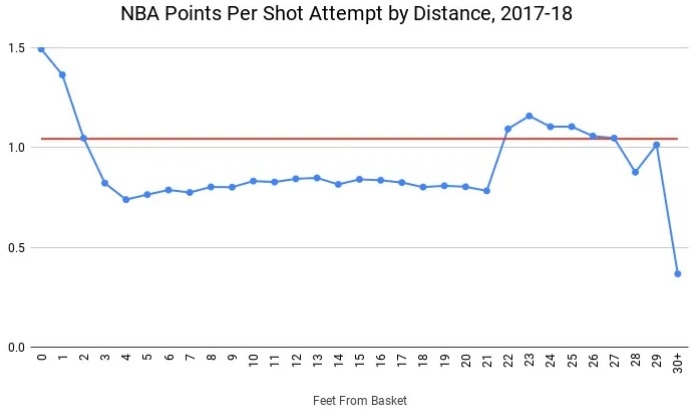


Figure 3: NBA points per shot attempt by distance, using tracking data from NBA Savant for 2017-18 season. Red line indicates expected point value across every attempt. (Kram 2019)

Similarly to the Oakland Athletics baseball team, the Houston Rockets have been seen as the trailblazers in this data driven tactic. Figure 4 shows that not only is the entire league attempting far more 3 point shots in the last decade, but the Houston Rockets are a team that believe in this statistic so vehemently that they are now electing to shoot

over 50% of their shots from behind the 3-point line in nearly every single game they play. This is a dramatic display of the impact and influence modern sports analytics has had.

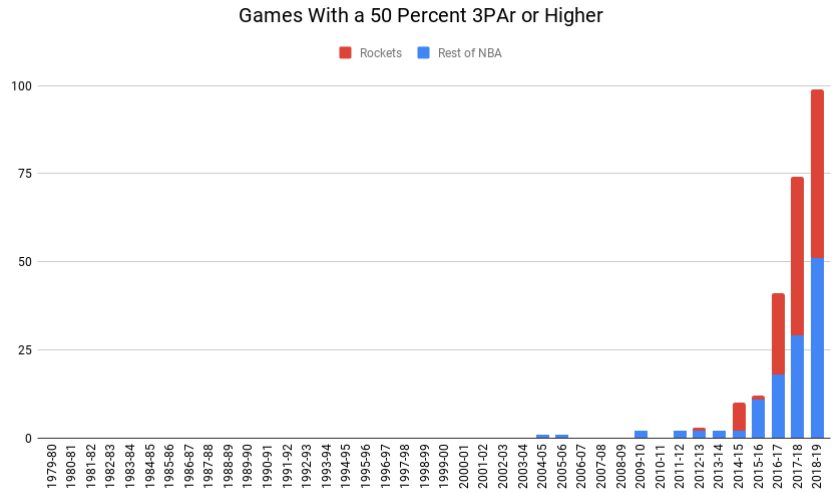


Figure 4: Number of NBA games with >50% 3 point shot attempt rates. Red represents the Houston Rockets and blue represents the rest of league average. (Kram 2019)

There is a limited amount of resources that provide empirical evidence of data analytics and technology having a profound effect within the sport of football. This is most likely due to a combination of football being earlier in it's life cycle of high-level data analytics being incorporated into the sport (compared to sports such as baseball and basketball) and football teams and organisations not wishing to publicise the practises they currently use in order to withhold a competitive advantage over their rivals. It is however widely accepted and acknowledged that football is currently undergoing fundamental changes across the board due to the influence of data analytics. There are few documented examples of long term success and impact via the inclusion of data analytics within decision making, particularly beyond player scouting and external analysis such as journalistic endeavours. There are, however, a small number of elite level teams that have found success through the use of data analytics although the effects of this are yet to be quantified. Brentford FC are a notable club in this endeavor. Over the last 15 years Brentford have seen a dramatic overhaul of staff and operations that has seen an acquisition of mathematicians and analysts in order to, like Oakland Athletics, find 'hidden gems' that are performing to an elite level in statistics (such as Expected Goals) that other clubs may not be paying close attention to. Since this dramatic restructuring at Brentford, culminating in a promotion to the Premier League (the highest division in the English footballing pyramid and a league that is often described as the toughest and most competitive in the world), the clubs analytics based model now competes with some of the worlds biggest clubs, and best teams, and has since held it's own, finishing in 13^{th}

position out of a possible 20, securing their participation in the competition for another season.

Regardless of the lack of documentation of the long term effectiveness of sporting decisions being influenced by data, there is a general consensus among analysts that modelling football using existing data science techniques has the potential to be a powerful tool and there is still a large amount of progress to be made in the area.

3.1 The Expected Goals Model

In football, expected Goals (xG) is defined as the statistical likelihood that a shot results in a goal (Lucey et al. 2015). xG is perhaps the most prominent statistic to appear from the rise in interest in data science and analytics in football. There are many practical applications of xG. Sumpter suggests that when attempting to analyse a teams performance using xG (either their finishing abilities or their ability to generate shooting chances that can result in high xG shots in the first place) it is particularly useful over longer periods of time (7+ matches) (Sumpter 2017). xG is also used amongst analysts and scouts to evaluate if teams are scoring the amount of goals that they should be, and the potential reasons for this, such as not being in the correct positions where they are likely to score goals often. Analysts also use xG to assess team performances during matches in which the factor of luck may have been notable in the final scoreline. Player scouts use xG to evaluate the finishing capabilities of potential signings. For example, if a player's goals scored is consistently below their xG value, then they may be considered a poor finisher of chances, and vice versa. Or if a play-maker is consistently making passes that result in high xG shot chances for their team mates then they are most likely a strong decision maker. The xG statistic has become omnipresent in modern football analysis. It can often be seen in post match statistical breakdowns provided by mainstream media coverage providers such as Sky Sports, BBC and BT sport. It has even found its way into the popular video game franchises FIFA and Football Manager, the latter of which uses a feature called xG match stories, which suggests that the strength of xG is so great, it can be used to show the back and forth of dominance within a match (SIGames 2021).

Lucey et al. (Lucey et al. 2015) aimed to present a method of accurately estimating an xG value for specific moments in matches, by using a combination of strategic data (such as defender proximity, speed of play) and positional data. In football, there has always been an acknowledgement that there is a difference between shots in their likelihood of resulting in a goal, but prior to the research by Lucey et al. this value had never been quantified. Their study found that not only can xG be quantified, but done so to an extremely high level of accuracy with a minimal number of external factors beyond shot location, through the use of logistic progression modelling performed on previous data, a technique very similar to those done similar in one of the trailblazing sports of data

analytics, Basketball (Yue et al. 2014), (Miller et al. 2014), (Cervone et al. 2014). Lucey et al. established that the match context also plays an important part in determining the likelihood of a goal, and decided to divide shots into six different categories depending on their context: open-play (defined as possession in the forward third), counter attack (defined by a fast break by the opposition team from one end of the pitch to the other), corners, penalties, free kicks and finally shots that are the results of set pieces (e.g. a shot from a cross from a free kick or throw in) (Figure 5). This study has provided the framework for xG models since its publication, and a simplified version will be developed during this project.

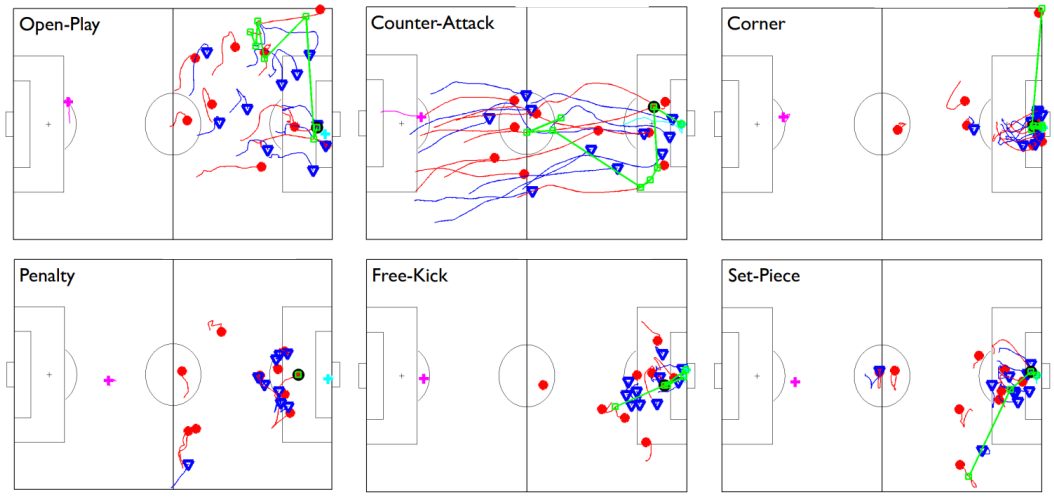


Figure 5: Example scenarios of the six different match contexts used to divide shots xG study by Lucey et al. (Lucey et al. 2015)

In defining xG and modelling it to the highest accuracy possible, there is an implication within the definition. Statistically speaking, if a team and/or players within a team attempt to play to maximise their potential xG over the course of a match or even season, then they stand the highest chance of scoring as many goals as is possible in any given match. This is one of the core implications of the xG statistic (and almost any expected value statistic within sport) and is a driving reason behind the inclusion of the model used for this project.

3.2 Expected Threat and Expected Possession Value

Expected Threat (xT) (sometimes referred to as Pass Value, Pass Impact, or even simply referred to as a Markov Model) is an attempt at quantifying how an action performed on a football pitch changes the probability of scoring. The resulting change in probability is the xT. xT is a metric used to assign a scored value to a defined number of locations on a football pitch based on the probability that having possession at that location will

result in a goal Figure 6.

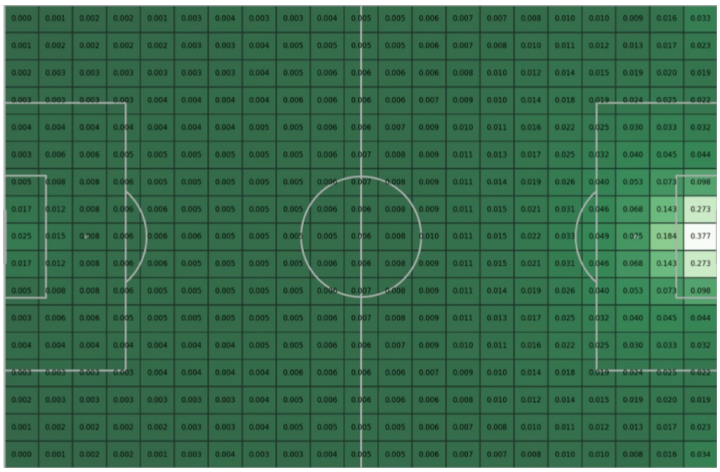


Figure 6: Expected Threat of different locations of a pitch. (Sumpter 2021)

Expected Possession Value (EPV) is another emerging framework that has been developed to evaluate the effectiveness of actions taking by a player in a match scenario, although this metric is designed to give a more holistic interpretation of the state of a match at any given time. EPV was developed to assign a numeric value (between -1, 1) to score every moment of ball possession to assess how likely a team is to score from that possession (-1 meaning the opposition team is more likely to turn over possession and score, 1 meaning the team in possession is likely to score) (Fernández et al. 2019). To do so Fernandez et al. predicted the likelihood of every possible the decision a player can make whilst in possession based on the spatio-temporal data surrounding that moment. This is done so using a Markov Chain model, where the states contained within the model are the possible actions the player can make. The actions were condensed into three categories: passing to a team-mate, shots on goal and holding the ball (dribbling to another area on the pitch). This model also takes into account the opponents surrounding the possession and the danger the pose to the attacking team.

3.3 Available Data

Despite the rising interest in football data, there still remains a lack of public accessible data to analyse, in academic and industry based contexts. Wyscout, one of the global leaders in football data collection, has released the largest publicly available footballing dataset to ever exist, for the purpose of aiding academic research and technological development (Pappalardo et al. 2019*b*). This is the dataset that was used throughout this project.

The collection contains all spatio-temporal events (such as shots, passes, tackles, etc.) that occurred throughout every match from and entire season of seven elite level football competitions: La Liga, Serie A, Premier League, Ligue 1, Bundesliga, UEFA European

Championships 2016 and FIFA World Cup 2018. Each event contains information about the time during the match which the event occurred, the two-dimensional location coordinates on the pitch, the outcome (i.e. did the shot event result in a goal/block/miss), the player(s) involved in the event and various other characteristics.

4 Methodology/Application

4.1 Application Design

The first step in developing the application was to create an initial design for the final outcome. As the application would be using data to generate a scoring system within the game, having a clear vision of the outcome of the application would allow the data and game-engine technology to operate in tandem throughout the project. The data modelling must return values that are applicable to the environment created in the game-engine, and vice versa. This process would also set reasonable boundaries and limitations to the development, ensuring the work undertaken was achievable within the available time frame.

The first stage of the design process was to assess the available data and decide which data to include in the project, and how that data was to be modelled. It was decided that the most useful (and relevant) data available from the dataset provided was the *Events* dataset. As mentioned, one of the key aspects of this dataset was that all recorded events within it had their 2-D location recorded and stored (start and end locations for specific event types such as passes). This meant that in order to combine this data with a game-engine environment, a two-dimensional display would be sufficient. This would also aim to simplify any interface the player has to interact with, and allow for fairly simplistic scenarios to be created that did not require an excess of tactical knowledge of the sport.

With this decision, the next step was to plan the potential environment in which the player will find themselves interacting. From the outset, there was a clear vision in the development of the application that it would involve a player being situated in a specific scenario. The player would then have to make a choice in the scenario based on the options presented to them (teammates available for passes, or shooting). Their choice would then be scored on the statistical analysis of the outcome of their choice resulting in success. It was decided that the chosen scenarios to be simulated within the application would be a simple passing scenario. The player will be randomly positioned at a point on the pitch, and a number of their teammates would randomly spawn at other locations on the playing field. The player would then have to simply select the team mate they wish to pass to.

For this game, an entire football pitch was used as the game environment. The pitch was split into a grid-like array of distinct zones which will be used for potential player

positions (Fig. 7).

The next stage was to decide which data within the dataset to focus on, and assess the way in which this was to be modelled, in order to obtain the necessary models to not only assess the decisions made by the player, but to do so in a form that was digestible and relevant to the sport. As per the above, both xT and EPV models are currently used by analysts to assess a players ability to progress the ball, and do so in a way in which the comparison of each model at one area of the pitch in any scenario differs from another. Because of this it was decided that a data model would be created that would build on the foundations of more complex models such as xT and EPV, but condense them in a manageable format from a data science point of view, whilst still providing the context and analysis relevant to ensure the integrity of the scoring system within the gameplay scenarios is kept. This model is a combination of the likelihood of the pass being chosen by the player resulting in success, and also the likelihood that the team-mate chosen will score (xG). This model will be defined and explained in more detail below. The key part of the model design was that the data would correspond to specific areas in the pitch, so that when player/teammate locations were randomised for each individual scenario, the corresponding data could be extracted and applied appropriately within the game-engine.

The final stage of the design process was to decide how the data would be displayed to the player. It was decided that the player would have their accumulating score visible to them throughout the gameplay loop to assess whether the decision they made has a high score (based on the PCR model) or not. Upon the completion of a set number of scenarios or 'rounds', the player would then be able to see a breakdown of the score given to them for the choice they made in each scenario, as well as the maximum attainable score for that scenario based on the statistical probability of success. An important design decision for the data display aspect of the project was the decision to allow the player to enter a post round review tool. This would allow the player to view back each scenario they have just played, to assess why their decision was or wasn't successful, and to receive a break down of the statistics that formed the scores for each available selection.

4.2 Creating the game environment

For this game, an entire 2D football pitch was used as the game environment. The pitch was drawn, and the scale and proportions of the pitch were compliant with international regulations for competitive matches (IFAB 2022). The pitch was split into 72 (12 x 6) distinct zones which will be used for potential player positions. As Wyscout uses coordinates of (0, 100) for both x and y axis' of the pitch, the x coordinates of the pitch were calculated by dividing the length of the pitch (100) by 12 and the y coordinates of the pitch were calculated by dividing the width of the pitch (100) by 6 and increasing in increments of the same value. As the midpoint of each zone would be used as the

player/ball spawn position the width/height of each zone divided by 2 was added to obtain the coordinates seen in Figure 7.

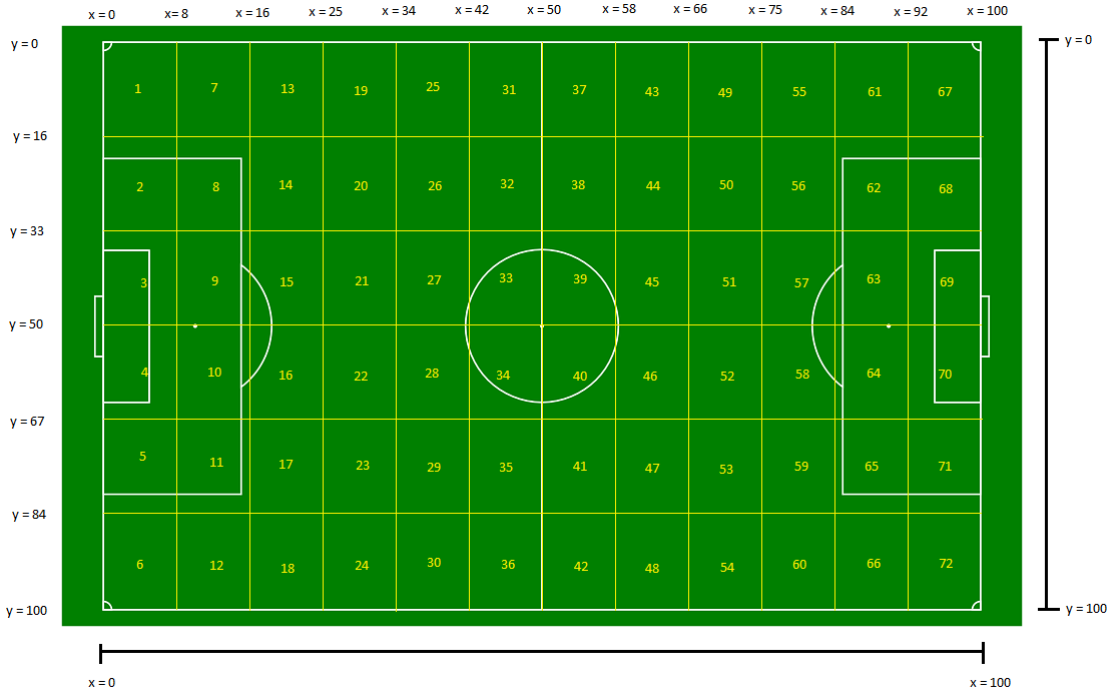


Figure 7: Diagram showing pitch layout used for in game zones and corresponding coordinates on pitch.

4.3 Creating a Pass Choice Rating Data Model

As mentioned above, a data model was created that combined the likelihood of passes succeeding from one area of a pitch to another, and the xG of the end location on the pitch. This model was inspired by the statistical output xT and EPV would return, but was designed in a way that would reduce the amount of data needed to create the model. This model will be referred to as the Pass Choice Rating (PCR). In order to calculate the PCR for any given passing scenario the percentage likelihood of a pass being made from the ball position to a team mates location is multiplied by the xG of the team-mate based on their location of the pitch. In order to build the PCR model, which will be used to score actions within the game environment and ultimately form the basis of our scoring system in game, both a model for xG and PSR were built. The formula used for the PCR model can be seen in Equation 1 where $pass\%_{a,b}$ is the likelihood of a pass being successful from zone a to zone b and xG_b is the xG values of a player located in zone b.

$$PCR = pass\%_{a,b} \times xG_b \tag{1}$$

All events data was stored in the JSON format and was extracted, manipulated and

modelled in the Python language.

4.3.1 Building an Expected Goals Model

The xG model built was an adaptation of a tutorial provided by David Sumpter at Friends of Tracking as part of the course *Mathematical Modelling of Football* (Sumpter 2022). The code used and adapted to build the xG model can be found on the Friends of Tracking GitHub page (of Tracking 2020). The xG model uses machine learning to fit a logistic curve to a combination of statistics such as distance from center of goal and angle between goal posts (Fig. 8) with respect to likelihood of scoring to present an accurate model of xG for any position on the pitch in order to reverse engineer the xG values needed to form our PCR model.

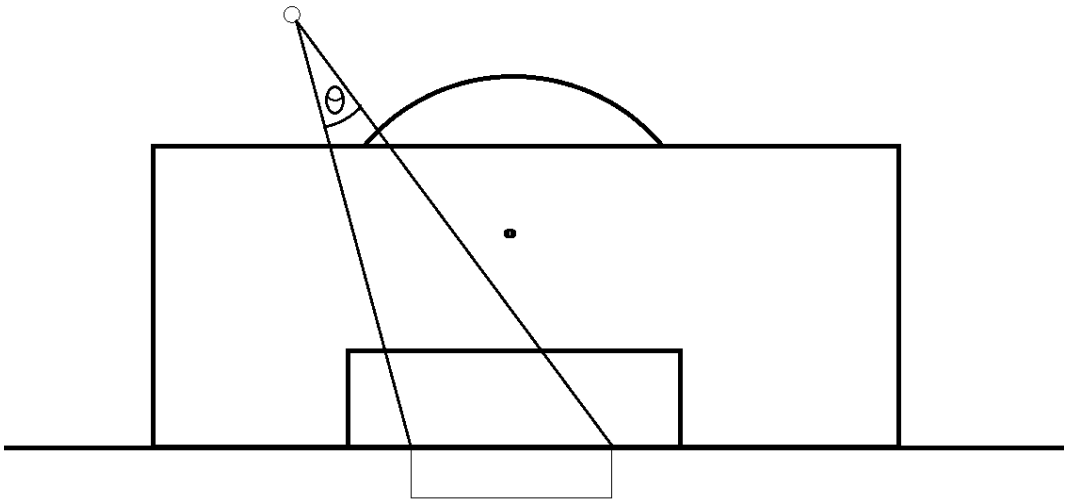


Figure 8: Diagram showing the being referred to as the 'angle between goal posts'. This angle is calculated using Equation 3.

The first stage of building our xG model was to extract data on shots taken from our datasets. For this project only non-headed shots were used. The data used was all shooting data from the top 5 league datasets plus the 2016 UEFA European Championship Data. Unfortunately due to an unknown data corruption error the World Cup data was not used. This data was read from the JSON filed and parsed into tabular data using *pandas.DataFrame* utilising the *pandas* library (pandas 2022) and stored alongside information on whether the shot resulted in a goal, and the x,y coordinates of the shot. The distance of each shot was also calculated (using basic geometry, see Equation 2), as was the angle of each shot (using Equation 3 which calculates the shooting angle between

both goal posts (Morales 2016)) and both were stored within our data frame, the reasons for this will be explained in more detail below.

$$d = \sqrt{x^2 + y^2} \tag{2}$$

$$\theta = tan^{-1}(\frac{7.32x}{x^2 + y^2 - (\frac{7.32}{2})^2}) \tag{3}$$

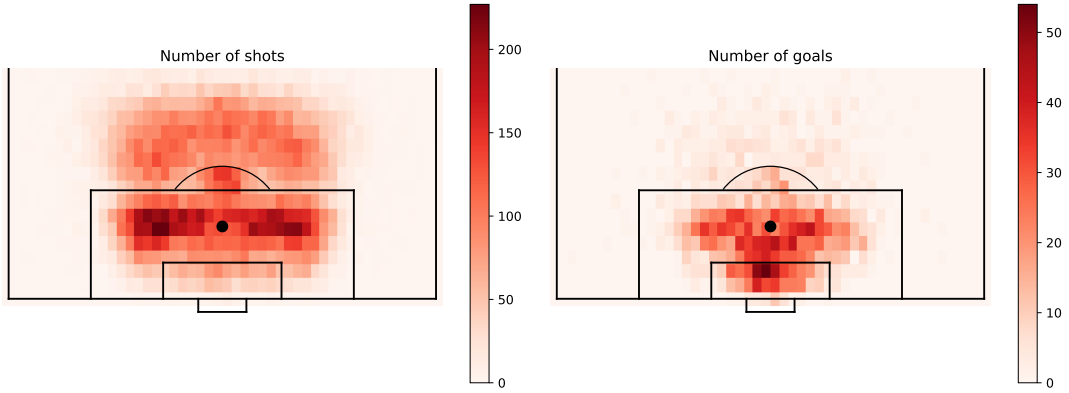


Figure 9: Heat-maps of number of shots (left) taken across competitions featured in Wyscout data and number of goals (right) as a result of these shots.

By comparing the number of shots taken by position to the number of goals scored from the same location we can build a very simplistic xG model based on the data available. A visualisation of this can be seen in Figure 10. This model would allow for the extraction of a purely statistical xG value for any location, based on the data available in the Wyscout datasets. Whilst this is technically a valid xG goals model, there are limitations of its practical use which resulted in the model not being sufficient for practical application within this project. Firstly as our game will feature coordinates from across the pitch (all the way back to attacking teams goal line) there would be lacking data the further from the goal we are positioned as there are very few (if any) shots ever taken from this range. The next problem is that even closer to the goal of the attacking team, there were locations on the pitch that over the course of the competitions used for data collection, there were only 1 shot. This meant that in certain locations there were locations that had a 1.00 xG rating that meant according to the model you were 100% likely to score a goal from that location every time you shoot, even though the location is in a obscure position, an example of which can be seen on the dark areas on the left and right channels of the box seen in Fig. 10.

An effective way to avoid these problems and enhance the overall accuracy and usability of the xG model within our application was to use machine learning and logistic regression tools to fit the data to a logistic model. The aim was for this model to be purely dependent on the x,y coordinates of the shot location as this was in line with the

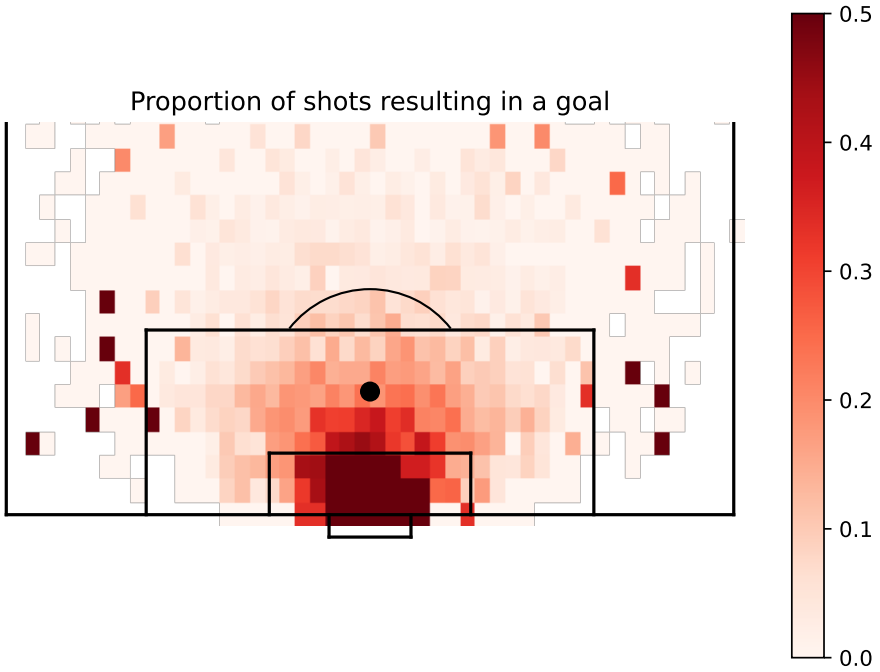


Figure 10: Heat-maps showing a purely statistical xG model based on the data available through the Wyscout dataset.

information that a player would have within the application. In order to model xG as a function of both distance from the goal and angle, the statsmodels library was used (statsmodels 2022) to find the best fitting generalised linear model. Using the statsmodels.formula.api.glm().fit() function, whose parameters contain the dataset we wish to find the fit for (in our case, likelihood of goals as a function of angle and distance) returns the parameters of our logistic function, which in the case of a multi variable logistic function would correspond to b_0, b_1 and b_2. These parameters can then be returned into our logistic function 4, where now to calculate the xG at any location we simply need to input the shot angle and distance of the shot based on x,y coordinates. A visual example of xG modelled using logistic regression can be seen in Fig. (note the levels of symmetry compared to fig. 10).

$$xG = \frac{1}{1 + e^{b_0 + b_1\theta + b_2 d}} \tag{4}$$

The coordinates of in-game positions of player's spawns were fed into the xG model created and the results were stored in an array of size 72 (corresponding to the layout of the game environment shown in fig. 7) and written to a text file, ready to be used in our game-engine, as shown in Appendix A.

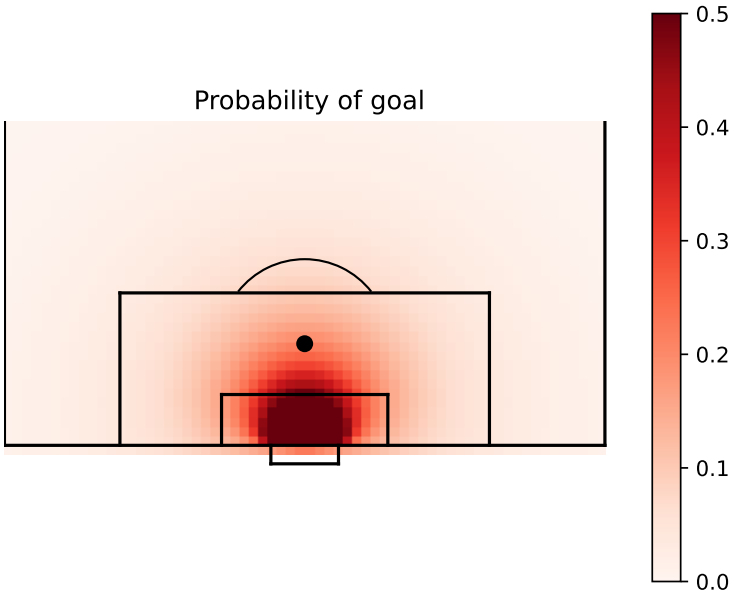


Figure 11: Heat-maps showing an xG model built using logistic regression on the data available through the Wyscout dataset.

4.3.2 Building a Pass Success Rate Model

A Pass Success Rate (PSR) model was also built using Wyscout events data and modelled in python. A simple model was built, as the dataset was vast enough to provide an accurate model for the purposes of this project. The first step of building the PSR model was to loop through each 'pass' event from the dataset, and obtain the start and end locations of the pass. Two dimensional arrays of size 72×72 were used to store passing data where the first iterator if the array would reference the starting location of the pass, and the second would correspond to the end location, where again the locations were grouped into their correspondence to the game-environment layout shown in Fig. 7. Three two-dimensional numpy arrays were used in the process, the first to track the number of passes made from one pitch zone to all other pitch zones and one to track the number of successful passes from each pitch zone to all pitch zones. From this the likelihood of a pass being successful from any zone to another was calculated by dividing the number of pass attempts for one zone to another by the number of successful passes between the same zones. This calculation was performed for each zone and multiplied by 100 to obtain a pass percentage success likelihood and these values were stored in a third and final two dimensional array. This process was performed for all six Wyscout datasets and the mean result for each zone to zone pass percentage was taken, resulting in our PSR model. Meaning that in order to access the likelihood of a pass succeeding from zone 1 to zone 2 on our pitch layout, for example, we would simply have to access the pass_percent array at index [0][1]. This array was output in the format of a text file, ready to be used within the game-engine. An example of how this data is formatted can

be seen in Table 1 which shows the percentage likelihoods of passes succeeding across the first 6 zones, where the second number on the first row refers to the likelihood of success from passing from zone 1 to zone 2, etc.

39	94	98	100	100	100	...
83	65	97	100	100	50	...
82	82	75	85	99	94	...
88	93	75	79	81	91	...
73	86	96	98	65	82	...
0	0	97	100	95	38	...
...	...	...	...	...	...	...

Table 1: Percentage likelihood of passes succeeding across the first 6 zones. For example, the second number on the first row corresponds to the likelihood of success of passing from zone 1 to zone 2.

With both the xG model and our PSR model, it is possible to build our PCR model by simply assessing the start and end locations of any random scenario created within our game-engine, and extracting the pass % value from the possible start and end zones within the scenario, as well as the xG value of the possible end zones within the scenario, and inputting the values into the PCR equation described in Equation 1.

4.4 Unity Implementation and Game Mechanics

The following sub-section outlines the mechanics implemented within the Unity engine to build the game-environment use in the application and combine it with the modelled data described above. The techniques used a combination of unity software tools with scripting written in the C# language. The application has two main portions: the gameplay loop, in which the player is presented with a set number of scenarios in which their goal is to choose the optimal pass to a team mate based on maximising PCR score, and the review stage of the game, which allows the player to view which rounds they performed sub-optimally, and return to those scenarios and examine which choices would have resulted in a higher PCR score, by presenting to them the corresponding xG and pass % data used to calculate the scoring value of their choice.

4.4.1 Data Initialisation and In-Game Progress Tracking

The modelled data for both xG and PSR was loaded into the game through the use of the System.IO framework. A 2D array (passing_data) of size 72×72 was initialised to store the passing data and 1D array (xG) of size 72 was initialised to store the xG data. As described above, the first index of the passing_data array would correspond to the zone a pass originates from, and the second index corresponds to the zone the pass ends at. Similarly the index of the xG array refers to zone of which we want to obtain the xG

at. The code used to load data from text files into the Unity framework can be found in Appendix A.

A number of variables were created to track and store information about a players progress throughout the gameplay loop. Upon the start of a new game, the modelled data is loaded and stored into the unity framework. The players round count is initialised and stored as zero to track the number of rounds completed. The players score is initialised and set to zero to track their accumulative score as they progress through rounds, and a variable named maxScore is created which will store the maximum available score of all choices available to the player in any round. This variable will be stored per round and used later within the application to display to the player how much score has been lost through sub-optimal decision making, or whether or not they made an optimal decision.

4.4.2 Spawn System

The first core mechanic to implement within the game-engine was the spawning of teammates and player based on the 2D coordinates describe in the above sections. A pitch design was created and input into a Unity project, forming the environment for this game. The world coordinates for the 72 available spawning zones within the game environment were determined by evaluating the number of pixels used by the playable area of the pitch and dividing this number by 12 for the length and 6 for the height. With the coordinates of the center of zones determined, an algorithm was written that would loop through each object that was to be spawned (the player, represented by a ball, and their teammates) and select a random number between 0 and 71 (0 and 71 was used instead of 1 and 72 as arrays were used to store the coordinates that corresponded to the zones, and the index for these arrays begins at zero) that would correspond to a potential zone to spawn. A check was made after each random zone selection to ensure that a game object does not already exist in this zone. If the zone is empty, the position of the game object is updated and the algorithm will continue until all game objects have been positioned. When an object is placed, their position is stored in a 2D array named *PositionHistory*, in which the first index corresponds the current round being played by the player, and the second index corresponds to the game object being positioned. For example storing the position of the second game object in the third round would involve change the PositionHistory array at index [2][1]. This spawn system was used every time a new round was launched, either through the starting of a new game, or the progression from one round to another. An example of a random set of player spawns can be seen in Figure. 12. The random spawning of game objects was performed in the RandomSpawns() function.

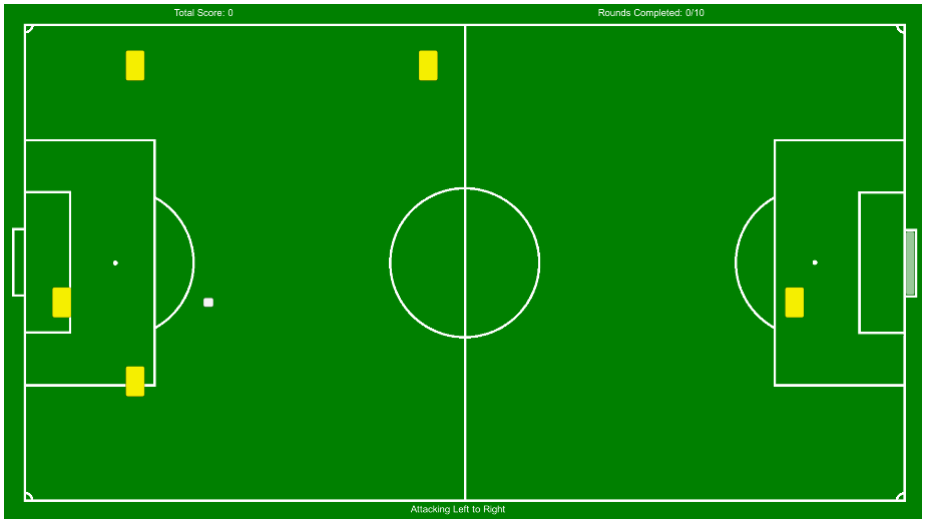


Figure 12: Example of random spawns of player and teammates upon starting a new game.

4.4.3 Gameplay

The game objects used to represent teammates within the game-engine were interactive UI buttons within the unity framework. This was done so that when a player wishes to make a choice of teammate to pass to, they can simply click the button represent the desired choice. Upon making their choice, the players choice of teammate is recorded and stored for use in the review portion of the game. At this point, a function is called that calculates the scores available for each teammate surrounding the player. This is done by cycling through each teammate, obtaining the xG of that teammate, and multiplying this by the likelihood of a pass succeeding based on the position of both the player and the teammate being assessed, this number was then multiplied by 100 to make the scores within the game easier to read for the player. The xG, pass % score and PCR scores are then stored in arrays, again for access and use in the review stage of the application. The players score is then updated based on the choice of pass they made by adding the score of their decision to their previous score. At this point, the RandomSpawns() function is again called, and the round count is updated, beginning the next round for the player with a new selection of passing options available for the player to assess and choose.

The player can continue to cycle through scenarios, accumulating score for a set number of rounds. When the maximum round count has been completed, all current game objects are removed (the ball and teammates) and the player is progressed into the review stage of the application, in which they can assess their performance throughout the rounds recently completed.

4.4.4 Performance Review and Scenario Evaluation

Upon the completion of the set number of rounds of gameplay, the player is presented with a review menu. In this menu the players score recorded at each round is presented to them, along with the maximum possible score they could achieve in each round. The player is presented with a menu of options, allowing them to either review each previous scenario, start a new game, or quit the game Fig. 21. This was achieved through a combination of Unity UI elements such as buttons and text boxes. If the player wishes to start a new game, the unity scene is reloaded, and all scores and information stored in player/teammate positioning from previous scenarios is reset. As the positions in each scenario are random, this should result in the player having a completely new set of scenarios to complete.

If the player wishes to review previous scenarios, they will be returned to the exact scenario they were presented with in the first round of gameplay. The choice they made in that round will be displayed to them by showing the chosen teammate as green, instead of yellow. If a player chooses to interact with a teammate in the review stage, this will toggle the display of relevant data associated with that choice. This data includes the xG of their teammate, the likelihood of the pass succeeding, and the PCR score Fig. 13. The players score achieved in this round is also displayed to them, and the current round they are reviewing is displayed. When the player has finished reviewing the scores and data available for each round, there is another button available to them that will progress the review stage to the next round. Once the player has completed their review of all rounds they are again presented with the review menu in which they can return to reviewing scenarios, or again start a new game.

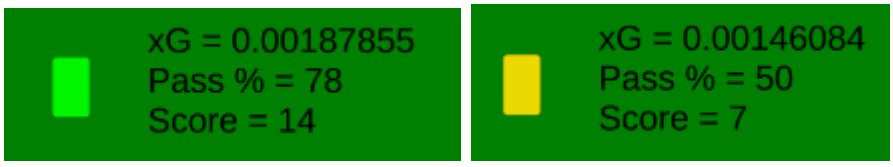


Figure 13: Example of change of display for choice made in review stage with stats displayed (left) and example of stats display for teammates who were not chosen by the player (right).

5 Results and Discussion

The results of this application development will be assessed on the standard to which the technology developed can be considered a useful tool in it's proximity to the purposes outlined in the research question and the uses discussed throughout this paper.

The results of this project will be assessed in three sections of development:

1. Modelling Data that is accurate and relevant to the gameplay aspects of the application. This data must also be usable within the game-engine and presentable to the extent that players can assess their performance in game when presented with this information.

2. The overall design of the main gameplay loop within the application. Does the application create an interactive game environment which combines the modelled data with a re-playable gameplay loop, to assess the quality of decisions made by a player.

3. Does the application provide the player with learning tools necessary to allow for analysis of their performance, whilst informing the player of the relevant data tools used in order to score their decisions.

5.1 Data Modelling

5.1.1 xG Model Results

Figures 14, 15 and 16 show the xG values found across the simulated game environment using our xG model, at varying levels of precision for the xG. 72 xG values were calculated using the centre coordinates of each zone, calculating the angle between the goal posts at this location, and the distance from the goal. These values were modelled as a function of likelihood of scoring, using logistic regression. Our final logistic model of xG can be seen in Equation 5 where θ is the angle between the goal posts (Fig. 8, Equation 3) and d is the distance from goal of the shot.

$$xG = \frac{1}{1 + e^{1.1979 - 1.6847\theta + 0.1043d}} \tag{5}$$

Figure 14 shows that the xG value grows according to the proximity of the player to the oppositions goal. Figure 15 shows the importance of progressing the ball towards the box if the desire is to maximise the xG. Figure16 shows that having the ball in the oppositions half will always result in an xG than if the ball remains in the attacking teams half. This provides an accurate representation to the player of the importance of xG within football analytics, following the assumption that progressing the ball to an area of the pitch that maximises xG will result in more actual goals over an extended period of time, or in the case of our application, will result in higher scores for their chosen action. The consistency of this result for the length of the pitch results in data that is reliable for the player, and when taken into account in the PCR model, should result in encouraging the player to progress the ball into 'dangerous' areas of the pitch in order to increase their score, regardless of their area of the pitch. This is something that would not be possible by simply using the raw data from the Wyscout dataset, as this

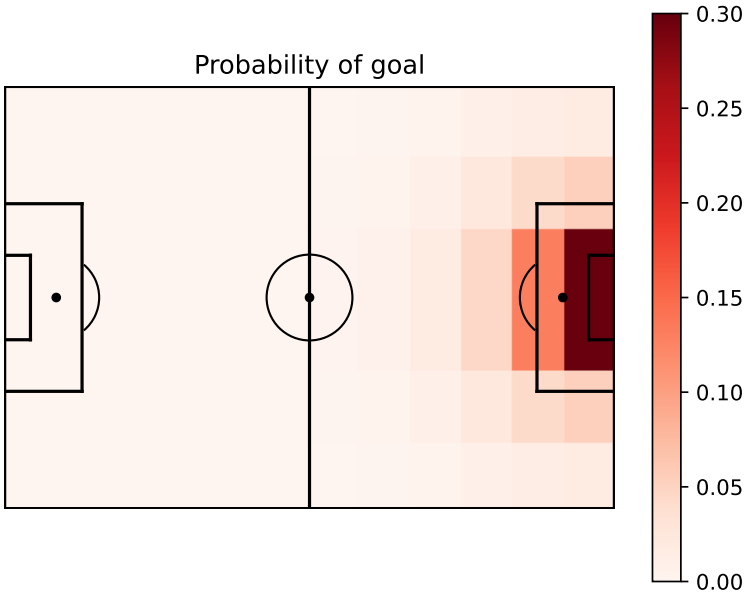


Figure 14: Heat-mapped results of xG for coordinates used in game-engine for a range of xG = 0 to 0.3. Attacking direction is left to right.

would have required data on shots being taken from every location on a pitch, when in reality it is highly unlikely that there would be a reasonable dataset of shots taken from inside the attacking teams own box.

A notable result of the xG model is the symmetry and consistency of the model in all areas of the pitch. If we compare the oppositions goal areas seen in Figures 10 and 11 we can see the difference in symmetry. This was a strongly desired result and a main motivation for attempting to model xG using logistic regression, as opposed to using a purely statistical model such as the one used to calculate the passing likelihood percentage. This was important as it would allow for consistent results within the game-engine to limit the amount the player would be relying on luck and/or memory for their performance scores, as opposed to contextualising the score calculations in a way that corresponds to genuine football tactics. If one of our xG models used in the game was a statistical anomaly from the data (i.e. a data point from a large distance and/or tight angle that had a supposed xG of 1.00) then this would be inconsistent for the player and their ability to make choices within the game-engine that appear consistent.

Although the results of the xG model may or may not have been accurate in modelling xG in real games (due to the limited number of variables used to create the model), the purpose of the model in this project was to form part of the scoring system used within the game-engine. The intention of this was to encourage the player to progress the ball towards to high xG locations through their pass choices within the game.

The range of values from the xG model were from $xG = 6.15 \times 10^{-6}$ for the zones closest to the attacking teams corner areas, to $xG = 0.312$ for the two zones closest to the

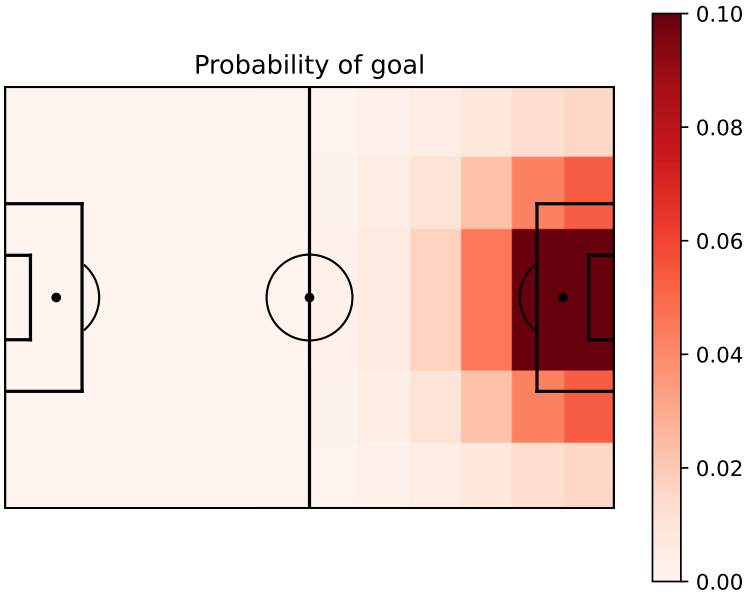


Figure 15: Heat-mapped results of xG for coordinates used in game-engine for a range of xG = 0 to 0.1. Attacking direction is left to right.

oppositions goal (A). These values differ by 5 orders of magnitude. As our PCR scoring value involves a multiplication of an integer value between (0, 100), a difference in order of magnitude of 5 can have a profound effect on the resulting score, again showing the impact of the xG model results on the scoring system within the game, and how this model may effect the decision making of a player of the application.

5.1.2 Pass Percentage Model Results

Figures 17, 18 and 19 show results for the pass likelihood percentage for passes originating in zone 26 (Fig. 17), zone 69 (Fig. 18) and zone 60 (Fig. 19) of our pitch layout (Fig. 7).

Zone 26 is displayed as it was the zone that contained the most passing events from our dataset, statistically returning the most complete model for passing percentage. The total number of passes performed from this zone in the dataset was $51,402$. Zone 69 is displayed as it was the zone that contained the least passing events from our dataset, with the total number of passes originating at this zone being 599. The difference in number of passes originating in zone 26 and zone 69 was ≈ 2 orders of magnitude. The mean number of inter-zone passes available from the dataset was $22,348$ and because of this, the results for zone 60 were also displayed, as the total number of inter zone passes originating from this zone was $22,525$, the closest value of all zones to the mean value. This inclusion of this result provides an example of a mean representation of the output of the model (or a close representation of what the average outcome looks like, based on the number of data points available).

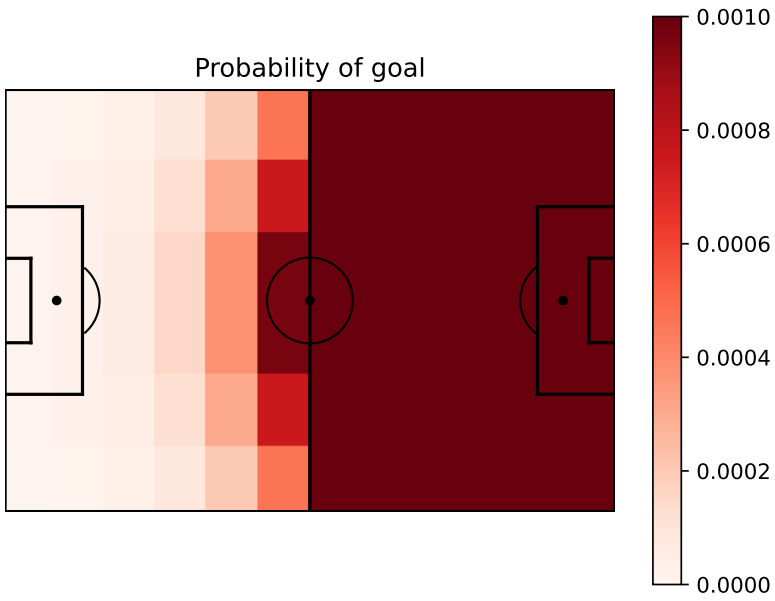


Figure 16: Heat-mapped results of xG for coordinates used in game-engine for a range of xG = 0 to 0.001. Attacking direction is left to right.

The pass percentage model had statistical shortcomings that may effect the overall application of the model within the game-engine. An example of a statistical shortcoming was that there were simply passing events from one zone to another that were simply never performed, and therefore did not exist in the dataset, resulting in the pass likelihood percentage for that pair of zones being zero. Another example of a shortcoming of the passing model was that there were cases in which a very small number of passes between a pair of zones from the dataset, resulting in pass percentage likelihoods of 100% for passes that may seem unlikely to a player within the application.

Figure 17 shows that for a large number of data points, a full and completed data model can be observed. The data model shows that the likelihood of passes succeeding into the oppositions box become increasingly unlikely. This may be due to the fact that the box is generally defended more resolutely than areas of the pitch. It can also be observed that are a limited amount of potential statistical shortcomings of this specific model, but they may exist. We can see that for passes originating in zone 26 and ending in zone 1 that the likelihood of this pass succeeding is 0%. It may be the case that this pass happened to be statistically unlikely to succeed during the competitions in which the datasets were obtained, but it is more likely that there was simply a very small number of events in which this passing scenario was observed. A passing percentage likelihood of 0% has an extremely profound effect within our PCR scoring system within the application as this choice of pass will automatically result in a score of zero, regardless of the xG associated with the end location.

Figure 19 shows that for the mean number of data points, a fairly complete model

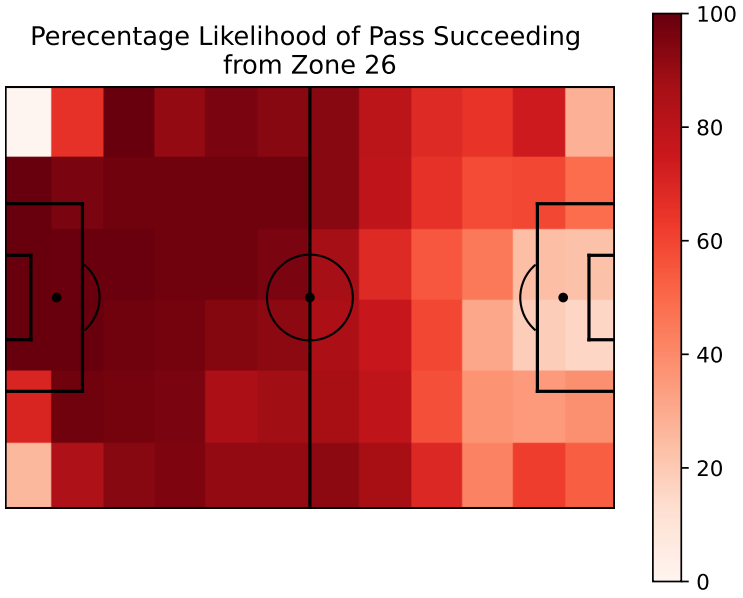


Figure 17: Heat-mapped results of Pass Percentage Likelihood for passes originating in zone 26 (Fig. 7) used in game-engine. Attacking direction is left to right.

can be observed. Again the model shows that the likelihood of passes succeeding into the oppositions box are unlikely compared to other areas on the pitch. We can also not that passes into the bottom right corner of the oppositions half are also somewhat unlikely ($> 50\%$). There are 11 zones across the pitch that have a 0% chance of passes originating from zone 60 succeeding. Once again, the reasons for this may that there is a small number of data points available from this point. However, we can observe that the 11 zones with 0% pass success is a greatly increased proportion of the pitch area when compared to Fig. 17. This result indicates that there may be a correlation between the reliability of this model and the number of data points used to formulate the model.

Figure 18 shows that for a low number of data points, the model appears very weak in it's practical application within the game-engine. A notable result is that in areas around the box, the percentage likelihood values are high. When incorporated into our PCR scoring mechanism, this would result in higher scores rewarded when the player keeps the ball in as close proximity to the oppositions goal as possible. 50 out of a possible 72 zones have a percentage likelihood of a pass succeeding of 0% from passes originating from zone 69. This result clearly highlights the potential drawbacks in the chosen method of modelling the passing percentage likelihood. As so many results are 0%, there is a high chance that our players team mates will be located in zones which have a percentage likelihood of 0%, and therefore in such an extreme case there is a significant amount of luck in ensuring that our player will have passes that give them meaningful and challenging decisions to make. There is a chance that if there are enough rounds in which the players starting positions are in zones which have a weak selection of passing

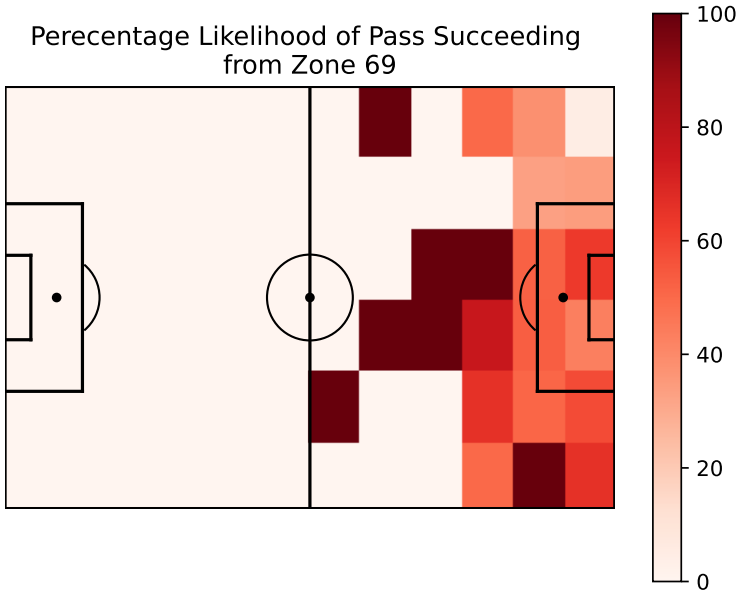


Figure 18: Heat-mapped results of Pass Percentage Likelihood for passes originating in zone 69 (Fig. 7) used in game-engine. Attacking direction is left to right.

data, then the player will not be able to assess their scores, and the potential scores of teammates in a meaningful way, as any player with a pass percentage likelihood of 0% will result in a PCR of 0. Having a large proportion of scores per round that are zero limits the amount of instruction the player receives from the statistics shown to them.

The results of the data model combine effectively to form the basis of a scoring system, albeit with some potential practical draw backs. Each scenario within the game-engine will have scores assigned to the actions contained within the scenarios, and each score used within the game will be formed from the data models built and designed using footballing data from real matches. The resulting scoring system used within the game is evidence that the combination of data analytics and game-engine technology was achieved.

5.2 Gameplay

5.2.1 Spawn-System

Figure 20 shows a full cycle of teammate and player spawns throughout an entire 10 rounds of gameplay within the application. The spawning algorithm effectively chooses player spawn locations and ensures that no two teammates are located in the same zone. There are times in which the player spawns do not replicate positions that may be expected in a real football match, as they are randomly selected. Examples of this can be seen in the fifth scenario of Fig. 20An effective way to improve the selection of player spawns may be to assess the passing data for available zones and only select locations that are within the top 10 most common passing pairs. Not only would this result in more

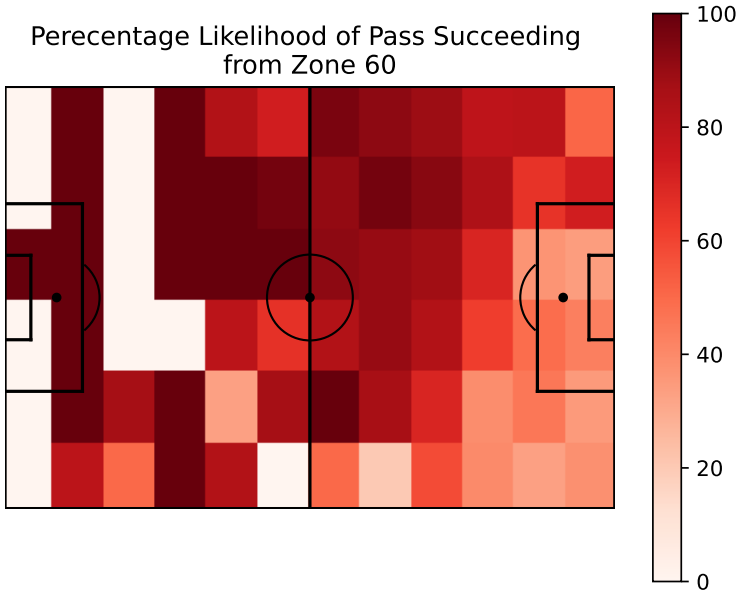


Figure 19: Heat-mapped results of Pass Percentage Likelihood for passes originating in zone 60 (Fig. 7) used in game-engine. Attacking direction is left to right.

realistic passing scenarios, but it would also counter some of the statistical shortcomings that are found within the passing data. For example if there is a dataset within the passing percentage likelihood model that has zero passes, this would never be selected as only the most common passing options would be selected. Another improvement on the spawning system would be to hand pick scenarios for the player to be in based, on real match scenarios. This would ensure that each scenario is grounded in realism and would provided an appropriate context for this purposes of this application.

The scenarios progress through each decision made by the player as intended, and when the set number of scenarios have been completed, the gameplay loop is broken, and the player is presented with a breakdown of their performance. As the player locations are random, and the gameplay loop re-playable, the player can continuously interact with the application until they are satisfied with their performance. The score calculations are performed within the game loop and effectively update and inform the player of their cumulative score as they progress through the game.

The integration of an infinite gameplay loop, with individual scenarios in which the player is faced with a simulated football action, where their performance is scored using relevant data analytics, is further evidence of the achieved combination of game-engine technology and football analytics. Once again the primary drawbacks in the results of this, are specifically related to the realism of the system in relation to real match contexts.

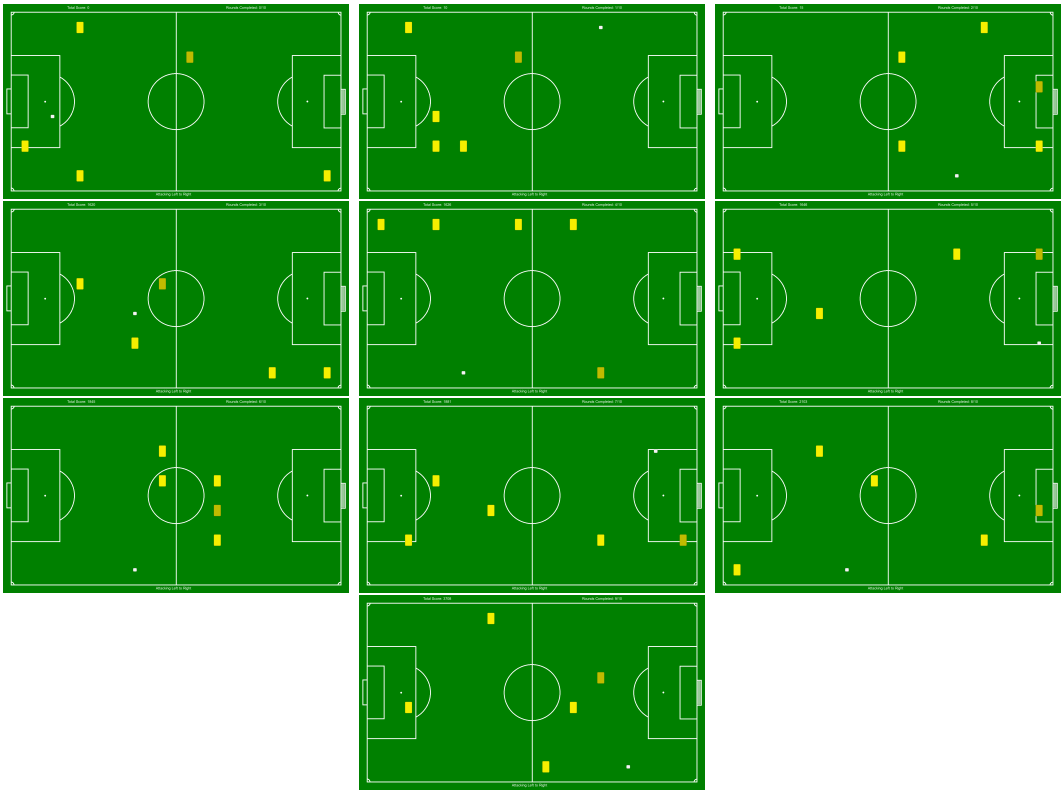


Figure 20: Random spawn locations of teammates throughout one set of ten rounds within the application.

5.3 Learning Tools and Review System

Figure 22 shows examples of the review system in practise. From these examples we see how a player of the application would be able to compare their overall score for a specific round with other choices that they could have made. The player is able to see the choice they made via the teammate being highlighted in green. The player is able to interact with each teammate and individually assess the statistics associated with that passing option. Each possible passing option is broken down statistically by the xG associated with that team mate, and the passing percentage likelihood of any passes to that team mate succeeding. The player can also see the total score available for each individual teammate, allowing them to assess whether the choice of pass they made was an optimal decision. The breakdown of statistics allows the player to explore the contributing factor as to why any decisions made by the player were suboptimal, with that aim that this will help guide players into understanding the data modelling that drives the scoring system within the game.

Figure 21 shows an example of the review menu that a player is presented with upon completion of a set of scenarios. In this interface the player is able to assess their performance by overall score per round, to give a general overview of their performance across the entire gameplay loop, whilst also allowing them the opportunity to identify the

Figure 21: An example of the review tool breakdown menu presented to the player on completion of the 10 randomly generated scenarios. This menu presents the player with a general breakdown of their performance, round by round, allowing them to analyse in which scenarios they achieved maximum scores etc.

rounds in which their choice was sub-optimal, so that when they review their performance within the review tool, they are aware of what rounds contributed the most to their score deficit.

The review tool combines the gameplay loop contained within the game-engine, and gives feedback to the player on their performance, whilst outlining the data analytical tools used to build the scoring mechanism and scenario design within the application. The incorporation of a tool within the application that allows the player to assess their performance within the application is evidence of the combination of game-engine technology, football data analytics and game-based learning being achieved.

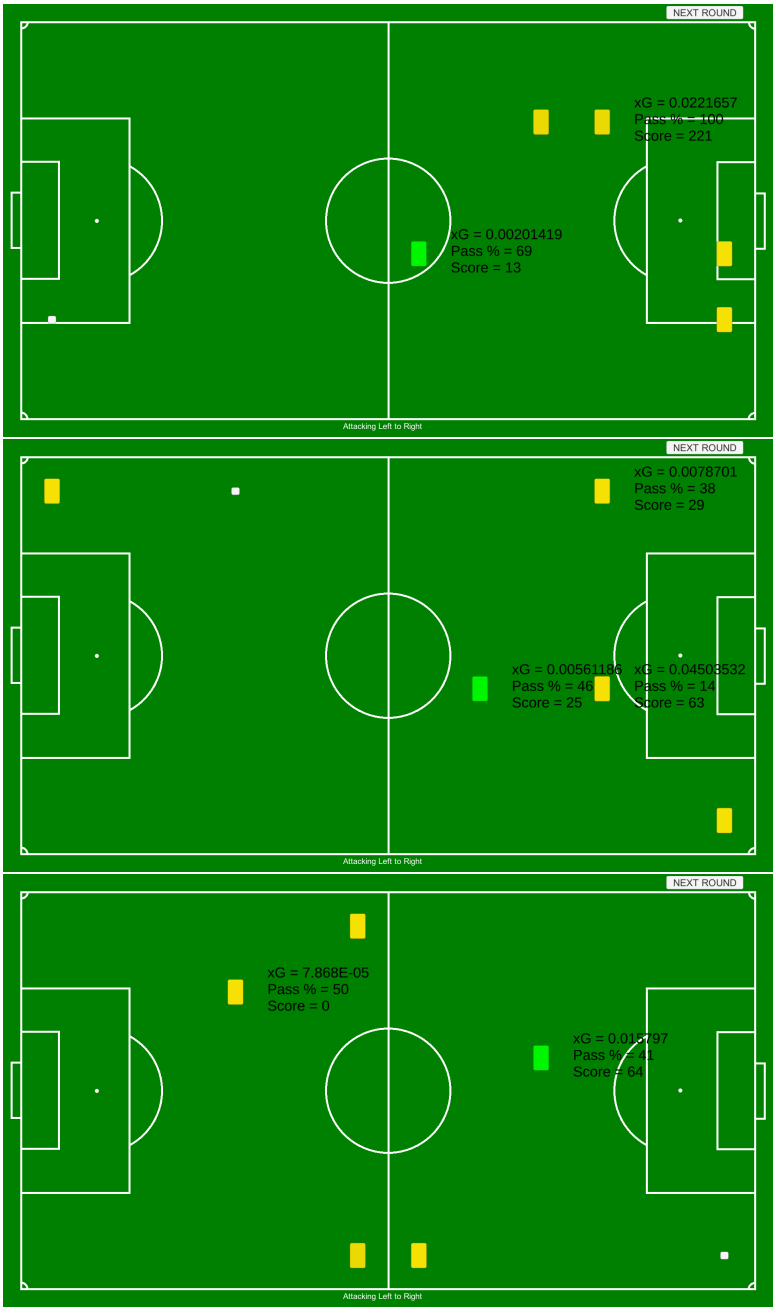


Figure 22: Examples of how the review system looks and works within the application. Each passing option is broken down by xG, Pass% and total available score, allowing players to revisit past scenarios and understand why their performance may have been sub-optimal.

6 Conclusion and Future Work

This project set out to investigate whether it was possible to combine football statistics and data modelling methods, with game-engine technology and game-based learning tools. The purpose of the developed application is to reinforce football tactics to users, and to provide a platform in which they can better understand data models that are currently used within in the sport. The player should gain an enhanced understanding of when and why these models are useful, and how they can inform tactical decisions made by the user.

In order to conduct this investigation, game-based learning and gamification tools were reviewed, to aid in the design process of the application. A scoring system and a review tool were identified as key methods of providing the platform for players to learn within the application. These were implemented via a combination of data models and relevant statistics used within elite level modern football, as well as a breakdown and display of the relevant data forming the motivations within the decision making aspects of the application. An application was developed in which the players finds themselves in a simulated football match scenario, where they have five possible passing options to teammates available, across the length of the pitch.

A scoring system was developed that combines modelled data from real football matches, and applying the result in a relevant and meaningful context. The aim of this was to provide the user with a motivation when making their decisions to always select the choice of pass which maximises the score available to them. Data models were chosen carefully so that higher outputs of the model corresponded to increased success in the actions being modelled. The scoring system also uses data models and modelling techniques that are currently used at a professional level within the sport, as well as a dataset containing statistics from elite level competitions.

An xG model was built using logistic regression on data gathered from real football matches, which returns the likelihood of a player scoring a goal from any specified area on a football pitch. Shot location in terms of angle from goal posts and distance from goal were modelled as a function of likelihood of the shot resulting from a goal. The method of developing the xG model proved successful for the purposes of this application. A pass percentage likelihood model was developed by assessing over $1,600,000$ passing actions performed across elite level competitions to construct a model that returns the likelihood of pass succeeding from one defined zone of a football pitch to another. Although the pass percentage likelihood model was effective enough for the purposes in this application, there is potential that this model in particular would restrict any practical applications of this research, as there are many data points within the model in which the results would be considered either unreliable or unhelpful. A possible remedy for this apparent drawback would be to perform 72 logistic regression models for each zone of the pitch.

For each model the position of start location and end location would be modelled as a function of likelihood of a pass succeeding. Doing so would provide consistent data to be used within the application, similar to the results of the xG data model developed.

Our two data models were combined to create a Pass Choice Rating which formed the scoring system within the application and is used to assess user decisions in each scenario. This point scoring system and the corresponding data models were designed and developed to provide an example of how the two technologies (football data modelling and game-engine technology) could be combined practically and again potentially do not provide any practical application within the sport beyond this. This point will be developed further below. This point scoring system does however provide a useful platform on which the gameplay aspects can be designed to supplement a user's understanding of the data behind the model.

A review system was developed within the application that allows users to review the various choices they made within the application. The user is presented with a general breakdown of their performance based on the scores they received per scenario within the application. The user is then able to review each scenario individually to assess the options they were initially presented with, with each component of the scoring system displayed to them. The user can see the choice they made in each scenario. This system allows the user to identify which was the key component in their decision which reduced their score (or not) based on the data models that were used to build the scoring system.

The main conclusion of this project is that it has proved that football data analytics and game-based learning tools can be combined to design an application within a game-engine which has the potential to enhance a players tactical understanding of the statistics behind the design of the game, and the simulated environment within. By combining the development of a scoring system based on data models and game-like scenarios within a game-engine and supplemented with game-based learning tools such as performance review, it is possible to create an application which has the potential to be used practically to reinforce football tactics to the user.

6.1 Future Work

The success of formulating a method of application design in which football data analytics and game-based learning have converged within game-engine technology has unlocked a number of potential avenues to assess the practical application potential of the technology that has been developed.

6.1.1 Football Specific Work

There are a number of aspects of future work on this research that are specific to the sport of football that would aim to enhance the overall practical uses of the application.

Specifically with respect to developing the players understanding of the data analytics that form the basis of design and gameplay within the application.

In the application developed for this research, a single macro-scenario (performing a pass to a teammate) from a real match was selected, in which an infinite number of randomised micro-scenarios can be formed. In future work, a number of macro scenarios would be developed that extend data related to passing and shooting statistics, and would combine a number of other data models used at a professional level. An example of another possible scenario would be shot choice based on shot direction, height, shooting foot, etc.

A key draw back of this research in it's immediate practical application is that it was not developed by a coach or tactician and therefore the contextual knowledge and intuition on what is actually needed tactically from a player is non existent. Future work would revolve around the design of scenarios and data models used within the game-engine based on practical needs of coaches and teams professionally, or developing a framework that would allow for clubs to build their own scenarios based on their needs. Performing this is also limited by the data available for public use. Teams and coaches may realistically require data on future opponents and the data available in the Wyscout dataset only contains historical data and may not be relevant in some cases, but professional clubs often have access to vast datasets and in house data scientists to allay these constraints.

6.1.2 User Testing and Participation

In general, a notable aspect of future work that would be a main focus would be the incorporation and development of user testing within the application. A number of approaches could be made to incorporate user testing into future research.

In the future, performing an assessment of player performance within the application such as an Action-Cost analysis proposed by Kavanagh and Miller (Kavanagh & Miller 2021) could be particularly insightful. The purpose of this work would be to assess whether or not the tools afforded to the player within the application influence their decision making within the game and motivate them to minimise their proximity to optimal decisions made within the game-based on the scoring/winning mechanisms. If the data analytics forming the scoring mechanisms within scenarios were designed rigorously then this would result in a greater intrinsic understanding of this data for the player.

Another aspect of user participation within the future of this research would revolve around the design of the application and the data display within, based on the knowledge and experience of the models that the users within the test already have. Having a predesign analysis performed on the test group would allow for an enhanced level of understanding of the extent to which players understand the data being displayed, allow-

ing the data display and learning tools afforded to the users within the application to be adjusted based on this knowledge. With this in mind, performing an analysis at the pre-design stage would also provide a baseline for each user in their understanding of certain data models and their effects. Quantifying this data would allow for a direct comparison throughout a lengthened period of interaction with the application to better understand the effectiveness of the technology in the context of genuine tactical understanding of the data analytics in match contexts.

6.1.3 General Application and Game-Engine Developments

An interesting piece of future work in the general development of the application would be to compare the performance of users across 2D, 3D and Virtual Reality (VR) environments. As one of the main practical aims of such an application would be to influence the decisions players make in real match situations, assessing whether the dimensionality of the environment in which the simulated scenarios are contained, would be relevant. This would be done by simply comparing the performance of users across various versions of the applications, in which the design of the scenarios and the data used are identical, but the interface is different.

Another potential avenue of research with this technology would be to assess whether the general methodology used within this research can be translated across sports, and begin to formulate a general model for combining game-engine technology, game-based learning and data analytics as a learning tool.

A xG.txt

0.00000615
0.00000808
0.00000928
0.00000928
0.00000808
0.00000615
0.00001487
0.00002004
0.00002335
0.00002335
0.00002004
0.00001487
0.00003575
0.00004968
0.00005884
0.00005884
0.00004968
0.00003575
0.00008524
0.00012297
0.00014873
0.00014873
0.00012297
0.00008524
0.00020088
0.00030373
0.00037728
0.00037728
0.00030373
0.00020088
0.00046558
0.00074747
0.00096201
0.00096201
0.00074747
0.00046558
0.00105239

0.00182766
0.00247123
0.00247123
0.00182766
0.00105239
0.00228699
0.00441317
0.00641830
0.00641830
0.00441317
0.00228699
0.00465893
0.01037345
0.01695205
0.01695205
0.01037345
0.00465893
0.00851124
0.02287244
0.04596236
0.04596236
0.02287244
0.00851124
0.01299268
0.04273795
0.12893106
0.12893106
0.04273795
0.01299268
0.01527611
0.05331828
0.31237314
0.31237314
0.05331828
0.01527611

B LoadData()

```
void LoadData()
{
    string input1 =
        File.ReadAllText(@"c: \Assets\Data\combined_Data.txt");
    int i = 0, j = 0;
    foreach (var row in input1.Split('\n'))
    {
        j = 0;
        foreach (var col in row.Trim().Split(' '))
        {
            passing_data[i, j] = int.Parse(col.Trim());
            j++;
        }
        i++;
    }
    using (TextReader reader =
        File.OpenText(@"c: \Assets\Data\xG.txt"))
    {
        for (int x = 0; x < 72; x++)
        {
            xG[x] = double.Parse(reader.ReadLine());
            xG[x] = xG[x];
        }
    }
}
```

References

Bateson, P. (2005), 'The role of play in the evolution of great apes and humans', *The nature of play: Great apes and humans* **1324**, 167.

Breuer, J. (2018), You learn what you play: on the fundamental coupling of playing and learning in humans and digital games, *in* 'Evolutionary Psychology and Digital Games', Routledge, pp. 167–178.

Cervone, D., D'Amour, A., Bornn, L. & Goldsberry, K. (2014), Pointwise: Predicting points and valuing decisions in real time with nba optical tracking data, *in* 'Proceedings of the 8th MIT Sloan Sports Analytics Conference, Boston, MA, USA', Vol. 28, p. 3.

Deterding, S., Khaled, R., Nacke, L. E., Dixon, D. et al. (2011), Gamification: Toward a definition, *in* 'CHI 2011 gamification workshop proceedings', Vol. 12, Vancouver BC, Canada, pp. 1–79.

Faiella, F. & Ricciardi, M. (2015), 'Gamification and learning: a review of issues and research', *Journal of e-Learning and Knowledge Society* **11**(3).
URL: *https://www.learntechlib.org/p/151920*

Fernández, J., Bornn, L. & Cervone, D. (2019), Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer, *in* '13th MIT Sloan Sports Analytics Conference'.

FIFPro (2022), 'At the limit: Player workload in elite professional men's football.', `https://www.fifpro.org/media/bffctrd1/at-the-limit.pdf`. Accessed: 2022-06-29.

Gazzard, A. & Peacock, A. (2011), 'Repetition and ritual logic in video games', *Games and Culture* **6**(6), 499–512.

Goldsberry, K. (2019), *Sprawlball: A visual tour of the new era of the NBA*, Mariner Books.

Hamari, J., Koivisto, J. & Sarsa, H. (2014), Does gamification work?–a literature review of empirical studies on gamification, *in* '2014 47th Hawaii international conference on system sciences', Ieee, pp. 3025–3034.

Herberger, T. A. & Litke, C. (2021), 'The impact of big data and sports analytics on professional football: A systematic literature review', *Digitalization, Digital Transformation and Sustainability in the Global Economy* pp. 147–171.

IFAB (2022), 'Laws of the game', `https://downloads.theifab.com/downloads/laws-of-the-game-2022-23?l=en`. Accessed: 2022-08-29.

Kavanagh, W. & Miller, A. (2021), 'Gameplay analysis of multiplayer games with verified action-costs', *The Computer Games Journal* **10**, 89–110.
URL: *https://doi.org/10.1007/s40869-020-00121-5*

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`. Accessed: 2022-08-26.

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

Lucey, P., Bialkowski, A., Monfort, M., Carr, P. & Matthews, I. (2015), 'quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data'.

Miller, A., Bornn, L., Adams, R. & Goldsberry, K. (2014), Factorized point process intensities: A spatial analysis of professional basketball, *in* 'International conference on machine learning', PMLR, pp. 235–243.

Morales, C. A. (2016), 'A mathematics-based new penalty area in football: tackling diving', *Journal of sports sciences* **34**(24), 2233–2237.

of Sport, S. (2022), 'Physiology of football: profile of the game', `https://sportsscientists.com/2010/06/physiology-of-football-profile-of-the-game/`. Accessed: 2022-07-29.

of Tracking, F. (2020), 'Friends of tracking - soccermaticsforpython github', `https://github.com/Friends-of-Tracking-Data-FoTD/SoccermaticsForPython`. Accessed: 2022-08-29.

Pac-man (1980), Arcade [Game].

pandas (2022), 'pandas documentation', `https://pandas.pydata.org/docs/index.html`. Accessed: 2022-08-29.

Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D. & Giannotti, F. (2019*a*), 'A public data set of spatio-temporal match events in soccer competitions', *Scientific Data* **6**(1), 236. https://www.nature.com/articles/s41597-019-0247-7.pdf.
URL: *https://app.dimensions.ai/details/publication/pub.1122140166*

Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D. & Giannotti, F. (2019*b*), 'A public data set of spatio-temporal match events in soccer competitions', *Scientific data* **6**(1), 1–15.

Pho, A. & Dinscore, A. (2015), 'Game-based learning', *Tips and trends* .

Plass, J. L., Homer, B. D. & Kinzer, C. K. (2015), 'Foundations of game-based learning', *Educational Psychologist* **50**(4), 258–283.

Sailer, M. & Homner, L. (2020), 'The gamification of learning: A meta-analysis', *Educational Psychology Review* **32**(1), 77–112.

Seaborn, K. & Fels, D. I. (2015), 'Gamification in theory and action: A survey', *International Journal of Human-Computer Studies* **74**, 14–31.
URL: *https://www.sciencedirect.com/science/article/pii/S1071581914001256*

SIGames (2021), 'Data hub', `https://www.footballmanager.com/features/data-hub`. Accessed: 2022-08-26.

statsmodels (2022), 'statsmodels documentation', `https://www.statsmodels.org/dev/index.html`. Accessed: 2022-08-29.

Sumpter, D. (2017), 'Should you write about real goals or expected goals? a guide for journalists.', `https://soccermatics.medium.com/should-you-write-about-real-goals-or-expected-goals-a-guide-for-journalists-2cf0c7ec6bb6`. Accessed: 2022-08-26.

Sumpter, D. (2021), 'Explaining expected threat', `https://soccermatics.medium.com/explaining-expected-threat-cbc775d97935`. Accessed: 2022-08-26.

Sumpter, D. (2022), 'Mathematical modelling of football', `https://uppsala.instructure.com/courses/28112/pages/2-statistical-models-of-actions`. Accessed: 2022-08-29.

Tetris (1984), Electronika 60 [Game].

The Barça Innovation Hub Team, F. B. (2020), 'Gamification of barça's female team training sessions during confinement', `https://barcainnovationhub.com/gamification-of-barcas-female-team-training-sessions-during-confinement/`. Accessed: 2022-08-18.

Yue, Y., Lucey, P., Carr, P., Bialkowski, A. & Matthews, I. (2014), Learning fine-grained spatial models for dynamic sports play prediction, *in* '2014 IEEE international conference on data mining', IEEE, pp. 670–679.