NOVA IMS
Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

Bringing objectivity and predictability to one of the most diverse and opinionated sports in the world by leveraging data

David Silva de Sousa

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

2022

Bringing objectivity and predictability to one of the most diverse and opiniated sports in the world by leveraging data

David Silva de Sousa

MAA

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

BRINGING OBJECTIVITY AND PREDICTABILITY TO ONE OF THE MOST DIVERSE AND OPINIATED SPORTS IN THE WORLD BY LEVERAGING DATA

by

David Sousa

Internship report presented as partial requirement for obtaining the Master’s degree in Advanced Analytics

Co-Advisor: Frederico Jesus, PhD, NOVA Information Management School.

Co-Advisor: Sudarshan Gopaladesikan, Head of Sports Data Science, Sport Lisboa e Benfica.

February 2022

ACKNOWLEDGEMENTS

I would like to express my gratitude to Sudarshan Gopaladesikan, Bruno Lourenço, João Barroso, Manuel Veloso, and Miguel Ferreira from Sports Data Science team at Sport Lisboa e Benfica for providing daily support and knowledge sharing throughout this learning experience.

Also, my gratefulness to Co-Advisor, Prof. Frederico Jesus, for his teachings, understanding, and guidance along this journey.

Last but not least, I want to acknowledge my parents for their unconditional support and motivation, always pushing me to achieve more.

ABSTRACT

Football institutions look to gain a competitive edge on and off the pitch. In this sense, this report showcases two data-led performance analyses that provide objective and measureable insights to support decision-making within clubs. Firstly, Pass Sonar, an interactive tool that represents the frequency and distance of players' passes by direction in order to understand historical passing behaviors. Secondly, Valuing Actions by Estimating Probabilities (VAEP), a data-driven framework for assessing the value of individual technical actions performed by players during football matches. This framework values all offensive and defensive action types (e.g., passes, shots, clearances, and tackles) using event data whereas lacking teams' positional structure, environment around the ball, and disregarding tactical and physical contributions captured by tracking systems. On the other hand, this method makes it scalable for teams to use as tracking systems can be expensive. Notwithstanding, both data-centric solutions focus on player-level and team-level performance evaluation relative to technical aspects of the game across the English Premier League during 2019/20 season. Thus, characterizing teams' playing styles using Latent Dirichlet Allocation (LDA) algorithm includes a tactical element in future work.

KEYWORDS

Sports Analytics; Data Science; Probabilistic Classification; Valuing Actions; Representing Passes; Playing Styles; Event Data; English Premier League

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ADASYN	Adaptive Synthetic Sampling Method
ANN	Artificial Neural Network
AUC	Area Under the Receiver Operating Characteristic Curve
CRISP-DM	Cross Industry Standard Process for Data Mining
DM	Data Mining
DMAIC	Define, Measure, Analyze, Improve, and Control
EPL	English Premier League
EPV	Expected Possession Value
FNN	Feed-Forward Neural Networks
KNN	k-Nearest Neighbors
LDA	Latent Dirichlet Allocation
LR	Logistic Regression
ML	Machine Learning
MLP	Multilayer Perceptron
NLP	Natural Language Processing
RF	Random Forest
RFE	Recursive Feature Elimination
SEMMA	Sample, Explore, Modify, Model, and Assess
SMOTE	Synthetic Minority Oversampling Technique
UEFA	Union of European Football Associations
VAEP	Valuing Actions by Estimating Probabilities
xG	Expected Goals
XGBoost	eXtreme Gradient Boosting

“Without data, you’re just another person with an opinion.”
- W. Edwards Deming

1. INTRODUCTION

For the past years, the importance of data science has been growing across a range of areas within clubs. Technological advancements in football have seen a variety of data-related fields enter the game. Analogously to on-the-field competition, operations away from the pitch are equally competitive across several areas of football performance, while the use of vast amounts of data generated across football operations provides a competitive advantage for clubs.

In this sense, the main goal of this report is to bring objectivity and predictability to one of the most diverse and opinionated sports in the world by leveraging data. To do so, this report is distributed among two projects. Firstly, Pass Sonar intends to represent player passes by direction to perceive pass tendencies per player, visualize the impact of the substitutes in terms of game creation, study passes behavior over time, and analyze unsuccessful passing patterns.

Secondly, Valuing Actions by Estimating Probabilities (VAEP) framework intends to value players' on-ball actions. It assigns a value to each technical on-ball action in event data based on its impact on the game outcome by considering the action probability of creating and conceding a goalscoring opportunity in the near future while capturing the context in which the action occurred. The presented work exploits VAEP presented by Jan Van Haaren (2019), whereas the framework was implemented from scratch. The developed methodology studies goalscoring opportunities by including the Expected Goals (xG) concept into the definition of the target variables instead of being goal dependent, as goals are a very rare occasion of the game and directly dependent on the finishing capability of the attacker. Furthermore, under the imbalanced learning problem, several resampling techniques were exploited in addition to stratified K-fold cross-validation implemented during the tuning of the hyperparameters.

The developed projects under this report were accomplished during a 6-month internship at Sport Lisboa e Benfica. This report shows that data practices associated with daily decision-making enhance how football clubs play the modern game. In this concern, the presented work seeks to support day-to-day processes within Sport Lisboa e Benfica through data-centric solutions. Nonetheless, the presented results focus on English Premier League (EPL) event data during 2019/20 season powered by Opta.

However, analytics in sports has its owns challenges (Davenport, 2014). Most of the performance indicators attributed to players are empirically decided by coaches and, therefore, they vary from coach to coach (Hughes et al., 2012). Additionally, this report disregards tactical and physical contributions captured by tracking systems, which can serve as a strengthening element for future work.

1.1. Company Overview

The 6-month internship which supported the work under this report was accomplished at Sport Lisboa e Benfica, a Portuguese Football Club founded in 1904.

This remarkable club, located in Lisbon, holds a glorious history at the national level, competing in 86 editions of the Portuguese First Division. SL Benfica won 37 titles, representing the most national champion club and the fifth European club with the most league titles. In addition, SL Benfica embodies the highest trophied club in Portugal by winning 26 Taças de Portugal, 7 Taças da Liga and 8 Supertaças de Portugal.

Moreover, SL Benfica, also known as The Reds, is a prestigious club internationally, conquering the most desirable European Cup[1] twice (1961 and 1962) and achieving five finals as runners-up (1963, 1965, 1968, 1988, and 1990). In the most recent years, SL Benfica reached the UEFA Europa League[2] finals in 2013 and 2014. The reds are notorious for developing valuable teams across multiple sports, having among others modalities handball, volleyball, basketball, rugby, and table tennis for both male and female teams. Likewise, SL Benfica forms Olympic medallists. Namely, Pedro Pichardo (triple jump) and Fernando Pimenta (canoeing) reached gold and bronze medals, respectively, in Tokyo 2020.

In terms of forming valuable players, the academy of SL Benfica is recognized worldwide. The Reds won best academy in the world twice in the last seven years as per the Globe Soccer Awards. Additionally, out of seven editions of the UEFA Youth League[3], SL Benfica was able to reach three under-19 finals (2014, 2017, and 2020). Furthermore, concerning net transfers value, according to *transfermarkt*[4], SL Benfica holds the highest balance worldwide between income and expenditures made exclusively through transfers since 2008/09, mainly due to the sales of former academy players João Felix, Rúben Dias, and Renato Sanches reflecting the value of *Made in Seixal* brand.

This institution has the biggest fanbase in Portugal, characterized by 230 thousand supporters, the second-best worldwide, symbolizing how valued and supported the club is. Furthermore, this togetherness conduct is also present on the emblem through the Latin sentence *"e pluribus unum"* translated to *"out of many, one"* that incorporates a unified way of behaving among its associates. The Reds are known as *"Águias"* (the Eagles), Figure 1 an eagle named *Vitória* (Victory) freely flies over the crowd before each home match in Estádio da Luz, which was inaugurated in 1954 having a capacity for 120 thousand spectators. In 2003 it was demolished and rebuilt to receive the Euro the following year. Currently, *"a Catedral"* (the Cathedral) houses more than 65 thousand seats, the largest stadium in Portugal.

Figure 1 - SL Benfica emblem

[1] European Cup is a tournament open to the league champions of all UEFA member associations, as well as to the clubs finishing from second to fourth position in the strongest leagues. This tournament was rebranded in 1992 to UEFA Champions League.

[2] UEFA Europa League is the second-tier competition of European football, ranking below the UEFA Champions League and above the UEFA Europa Conference League.

[3] UEFA Youth League is contested by the under-19 teams of the clubs competing in the UEFA Champions League group stage, plus the domestic youth champions of the best-ranked national associations.

[4] Transfermarkt is a German-based website focused on valuing football players' market value.

1.2. General Objectives

A football game is described by a sequence of events on the pitch. Respecting the English Premier League (EPL) season 2019/20, approximately 1 700 events are produced throughout 90-minute matches, the equivalent to 646 000 data points in a season-long time window. Intrinsically, there is potential to convert these vast amounts of data into value to support decision-making processes within clubs.

In this sense, during the 6-month internship at Sport Lisboa e Benfica, the key challenge is putting the increasing amount of data available in the right context and drawing effective conclusions. To do so, it was applied data-driven methodologies to produce meaningful knowledge which supports day-to-day processes.

In concrete terms, this report is divided into two projects.

1. Firstly, Pass Sonar intends to represent player passes by direction to perceive pass tendencies per player, visualize the impact of the substitutes in terms of game creation, study passes behavior over time, and analyze unsuccessful passing patterns.
2. Secondly, Valuing Actions by Estimating Probabilities (VAEP) intends to value players' on-ball actions. It assigns a value to each on-ball action in event data based on its impact on the game outcome by taking into consideration the action probability of creating and conceding a goal action in the near future, while accounting for the context in which the action occurred, therefore, allowing to convert everything that happens in the beautiful game to a single currency.

Hence, the presented projects focus on the fields of performance evaluation as well as studying technical behavior by analyzing event data powered by Opta. Tactical and physical data captured across tracking systems (e.g., GPS) will not be featured in this study.

In addition, the approach under this report concentrates on identifying patterns and behaviors that occur during 90-minute matches throughout a season-long period. However, training processes reflect on game performances. Thus, this study can be complemented by understanding how players' training efforts affect their weekend performances.

Even though the accomplished work was developed under a Portuguese Football Club, the analyses presented in this report focuses on English Premier League data. In this way, the projects developed in this study were replicated to the most competitive league in Europe according to the prestigious UEFA Rankings[5], which serves as a validation methodology. The report centers on 20 Premier League teams across 38 game weeks, reflecting 380 games over the 2019/20 season.

[5] The club coefficients are based on the results of clubs competing in the five previous seasons of the UEFA Champions League, UEFA Europa League and UEFA Europa Conference League.

1.3. Specific Objectives

This section will cover an introductory overview of each project developed.

1.3.1. Pass Sonar (player-level analysis)

According to Opta, passes represent on average 60% of player actions during games. Hence, passes are characterized as the most frequent action in football. The pass purpose may differ according to game context, namely, retain possession, vary the center of the game, leave pressure moments, support direct play, unlock the opponent's defense or even create a goal opportunity.

The Pass Sonar project develops a dynamic and interactive visualization that illustrates player passes by direction throughout a football match. In concrete terms, each player is represented on the pitch by a polar-coordinate based bars, where the bar length indicates average pass distance while the color expresses pass angle frequency. Therefore, it is possible to perceive pass tendencies per player, visualize the impact of the substitutes in terms of game creation, study passes behavior over time, and analyze unsuccessful passing patterns.

Additionally, the visualization reacts according to user demands. In other words, the user has the capability to filter the visualization by competition, season, team, match, players, position, time played, and complete or incomplete passes. Consequently, the visualization automatically readapts without any additional costs or efforts.

As a result, Pass Sonar can serve for opponent analysis, through recognizing the opponent's game builder that characterizes a target to mitigate or identifying the defensive line member with the highest unsuccessful passing rate which represents a high pressure trigger. In the opposite direction, this tool can also be used for introspective analysis, for instance, acknowledging if the game builder was active in the game or even comprehending if players followed coach principles (e.g., wide full-backs playing interior passes).

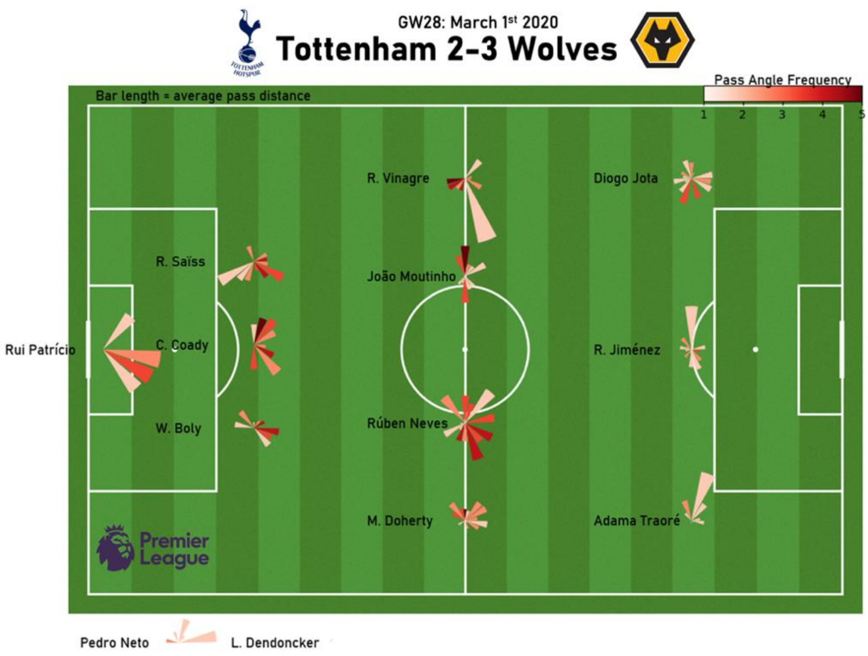


Figure 2 - Pass Sonar visualization

1.3.2. Valuing Actions by Estimating Probabilities (action-level analysis)

Theoretically, from a supporter point of view, it is intuitive to perceive which on-ball actions are valuable to the game. However, the challenge resides in developing a data-driven framework that assigns a value to each player's on-ball action.

Valuing Actions by Estimating Probabilities (VAEP) is a data-driven framework for evaluating player actions in football. It assigns a value to each on-ball action in event data based on its impact on the game outcome by taking into consideration the action probability of creating and conceding a goalscoring opportunity in the near future, while accounting for the context in which the action occurred (Decroos et al., 2019).

Notwithstanding, as goals are a particularly rare occasion of the game as well as directly dependent from the finishing capacity of the attacker, the developed framework predicts the likelihood of each on-ball action to create or concede a goalscoring opportunity, that is, a shot possessing an expected goal value (xG) above 20% or a goal. In this way, the sample size of the target variable minority class increases.

In this sense, an action value reflects the action's expected influence on the game, ranging from -1 to 1. Intuitively, positive values are associated with a contribution that, offensively, increases the team's likelihood of creating a goal action or, defensively, reduces the team's likelihood of conceding, whereas the magnitude represents the impact. As an example, an action valued at +0.10 is expected to contribute 0.10 goal opportunities in favor of the team performing the action. Contrarily, negative values are associated with negative contributions. That is, an action valued at -0.10 is expected to yield 0.10 goal opportunities for the opponent (see Figures 3 and 4).

The main advantage of this approach is that the framework considers all types of technical actions (among others, passes, crosses, dribbles, take-ons, shots, interceptions, and tackles). In the opposite direction, the current approach disregards off-ball tactical and physical contributions.

Figure 3 - Video representation of Bournemouth's goal

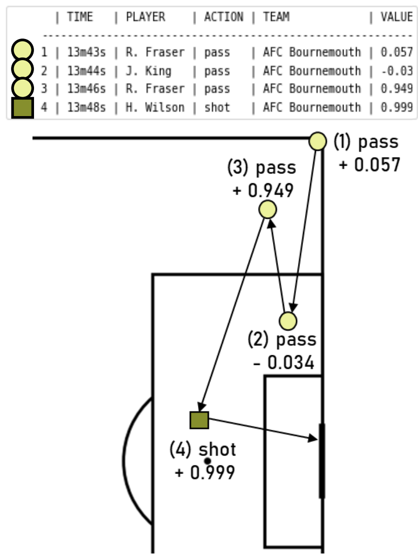


Figure 4 - Valuing on-ball actions

On this basis, the following practical/specific questions have been addressed:

PASS SONAR:

1) Who's the most and least involved player in terms of game creation?
2) Who's the game builder?
3) What's the impact of substitute players?
4) Who's the player that performs the most amount of unsuccessful passes?
5) Which player performs the furthest and shortest passes?
6) What's the time period where teams had more possession?
7) What's the time period in which teams performed more unsuccessful passes?
8) Does a player tend to pass forward, backward, or laterally?

VAEP:

9) How valuable is a player's on-ball action?
10) Who are the most impactful players in the league?
11) What's the performance characterization of a player over the whole season?
12) What's the performance characterization of a team on the entire season?
13) What are the teams performing more valuable on-ball actions in the league?
14) What are the most and least consistent teams in the league?
15) Is the result coherent with the performance?
16) How does a player perform compared to the remaining players in league with similar roles?
17) What's the most valuable position of a player?
18) What's the position that allows players to be more impactful?
19) What's worth more: high frequency or high value?
20) When playing simultaneously what's the most valuable duo or trio in the league?
21) What's the best replacement for a particular player?
22) What's the momentum of both teams during a 90-minute game?
23) What's the pitch zone where players perform the most valuable actions?
24) What's the most valuable pitch zone in which teams enter the attacking midfield, final third, or opponent's box?
25) What's the video footage mapping of players' most and least valuable actions?
26) Who's the most threatening set-piece team?

“Where there is data smoke, there is business fire.”
- Thomas Redman

2. THEORETICAL BACKGROUND

2.1. DATA SCIENCE

In today's data era, the amount of data created, consumed, and stored has grown exponentially (Van der Aalst, 2016). Intrinsically, this volume of data demands the need for someone to have the skill set and knowledge to transform this resource into valuable information, actionable insights, and wisdom to support decision-making within organizations.

The methodical systematic discovery of valuable relationships and patterns in data is enabled by a set of iterative activities collectively known as the data science process (Kotu & Deshpande, 2019). The standard data science process involves (1) understanding the problem, (2) preparing the data samples, (3) developing the model, (4) applying the model on a dataset to comprehend how the model may work in the real world, and (5) deploying and maintaining the models (Kotu & Deshpande, 2019).

From a non-technical perspective, this process of discovering knowledge from data includes rigorous study design, articulation of the objective or question, careful curation of data, assurance of data quality, an armament of advanced analytical tools, and steps to ensure reproducibility and reliability (Miller, 2020). To do so, data scientists require an integrated skill set spanning mathematics, machine learning, artificial intelligence, statistics, databases, and optimization, along with a deep understanding of the craft of problem formulation to engineer effective solutions (Dhar, 2013). Data is the scientist's intellectual and social currency, while the context of the data can be diverse, allowing data scientists to specialize in distinct industries such as sport, health, financial, insurance, automotive, retail, journalism, among many others.

The benefits of data-driven decision-making in organizations have been demonstrated conclusively. The Economist Erik Brynjolfsson from MIT and Penn's Wharton School conducted a study that analyzes how data-driven decision-making affects firms' performance. The study statistically reveals that data-driven decision-making is associated with higher productivity and market value and correlated with specific measures of profitability (Tambe et al., 2011), expressing the importance of successfully exploiting data.

In the coming decades, the scope and impact of this science will expand enormously as scientific data and data about science itself become ubiquitously available. In the future, open science will take over once sharing the underlying code and data will allow large numbers of datasets and analysis workflows to be derived from studies science-wide (Donoho, 2017).

2.2. DATA SCIENCE IN FOOTBALL

For the past years, the importance of data science has been growing across a range of areas within clubs. The use of vast amounts of data generated across football operations provides a competitive advantage for clubs. In this regard, the main data-driven analysis goal is to ensure that players, coaches, and support staff are equipped with the best possible information to make informed decisions and achieve day-to-day success.

Nowadays, game-related statistics in the sports industry are demanded by coaches, players, managers, journalists, supporters, fans, video games developers, betting markets, and academics (Zambom-Ferraresi et al., 2018). Besides football, such data analysis were developed in many other sports, including basketball (Cervone et al., 2014), american football (Romer, 2006), ice hockey (Routley & Schulte, 2015; Liu & Schulte, 2018), and rugby (Kempton et al., 2016).

In football, event data annotate a set of features describing specific discrete events typically involving the ball that occur during games. Additionally, tracking data records the location of players, referees, and the ball throughout games using tracking devices (e.g., camera-based or chip-based technologies). In this sense, the gathered data that supports football analysis may arise from internal sources, in which case, sports entities possess resources to capture data or external sources, in which case, data is obtained from outside the organization through specialized data providers (e.g., Opta, Wyscout, Stats Perform, SciSports and Second Spectrum).

Notwithstanding, data will not be replacing the human elements of performance and decision-making. Technical teams will always have their knowledge and judgment, while players will have their experience and instincts. However, data science can serve as a complement by the intelligent use of data to obtain actionable insights and provide tools to enrich decision-making. In this manner, the goal of data practices is to better structure and organize the thinking of coaches and players, not to replace them.

To do so, football data scientists work on all phases of the data lifecycle, facing complex projects and leveraging data engineering, data integration, data analysis, and data visualization skills to assist business needs. In addition, the developed work can be distributed among several fields, such as performance analysis, tactical analysis, opponent analysis, player development, player monitoring, fan engagement, financial analysis, and scouting. In these cases, both athletes and coaches benefit from the analysis.

Besides the internal work accomplished by data scientists at football institutions, third-party companies specialized in delivering data-centric solutions to enhance the way clubs and associations play the beautiful game enter the ecosystem of institutions.

The effects of institutions incorporating these researches into their ways of competing have had impacts on and off the playing field, namely, in terms of physiological enhancements of athletes, but also socio-political and economic outcomes (Thakkar & Shah, 2021).

2.3. FOOTBALL DATA SCIENCE STATE-OF-THE-ART

In order to comprehend the scientific background framing this work and adequately position its contributions, a review of the state-of-the-art in data science applied to football was undertaken.

Technological advancements in football have seen a variety of data-related fields enter the game. Analogously to on-the-field competition, operations away from the pitch are equally competitive across several areas of football performance. In this regard, data scientists and other personas of different non-footballing backgrounds have made the industry competitive and diverse.

Data analytics entered the sports field back in the early nineties. Ever since, everyone from individual athletes to major leagues has been using it in the service of performance, marketing, scouting, and other goals. Over the years, the intelligent usage of data applied to the football industry has grown as data availability increases. As a result, data practices associated with daily decision-making processes enhance how football clubs play the modern game.

According to Jan Van Haaren, Chief Analytics Officer at SciSports, the implementation of data analytics into professional football clubs can be divided into two categories. On the one hand, some clubs have invested in in-house data analysis or data analytics departments, while an ever-increasing number of clubs is leveraging data by collaborating with specialized companies that offer data-driven tools for day-to-day operations. Additionally, a hybrid approach can also be employed in some well-established clubs.

Hence, the key challenge for clubs is to put the increasing amount of data available in the proper context and draw the correct conclusions from it. In this sense, data scientists face the challenge of translating data into answers to the questions that practitioners (such as managers, coaches, sporting directors, and scouts) have in a non-technical language that they speak and understand.

In more concrete terms, data has been exploited across several business areas within clubs for player recruitment, injury detection, and prevention, opponent analysis, boosting team and player performances, increasing commercial efficiency, enhancing supporters' experience, social media engagement, among others (Davenport, 2014). Moreover, these analyses that extract helpful information, actionable insights, and knowledge from data can be executed pre-match, live, or post-match.

According to Oscar Ugaz, Former Digital Business Manager at Real Madrid, besides investing in software, clubs also need to invest in brainpower to analyze the captured data by hiring professionals capable of taking advantage of this valuable resource, reflecting the trident of people, processes, and technology.

Subsequently, practical approaches, frameworks, and metrics within the field of football analytics that impact the development of this report are described below:

- **Count-based approaches.** These techniques (McHale and Scarf, 2007; McHale, Scarf, and Folker, 2012; Pappalardo et al., 2019) rate players by (1) assigning a weight to each action type and (2) calculating a weighting sum of the number of times a player performs each action type (e.g., pass, dribble, cross, tackle) during a match. The weights are typically learned by training a model that correlates these counts with either the match outcome or the number of goals scored (or conceded).

- **Expected Possession Value (EPV).** EPV (Rudd, 2011; Mackay, 2017; Decroos et al., 2017; Yam, 2019; Singh, 2019) divides a match into possessions or phases, sequences of consecutive on-the-ball actions where the same team possesses the ball. Hence, these models value each action that progresses the ball, typically by seeing how much the action changed the team's chances of producing a goal-scoring attempt. Conceptually, most of these approaches can be seen as modelling a possession using a Markov model.
- **Valuing Actions by Estimating Probabilities (VAEP).** VAEP (Decroos et al., 2019) is an approach that goes beyond the possession-based ones by trying to value a broader set of actions and by taking action and game context into account. Decroos et al. frame the problem as a binary classification task and rate each action by estimating its effect on the short-term probabilities that a team will score or concede using boosting algorithms.
- **Expected Goals (xG).** Expected Goals (Pollard et al., 2004; Ryder 2004; Hamilton, 2009; Itjsma, 2011; Rudd, 2011; Green, 2012) measure chance quality. The objective is to calculate the likelihood of shots being scored from a particular position on the pitch during a specific phase of play by applying logistic regression analysis. This metric considers, among several features, the distance and angle from the goal, the space from the nearest opponent when the shot is taken, the body part used to perform the shot, and the run of play. As a result, each shot has an expected goals value which is the percentage change resulting in a goal.
- **Pitch Control.** Pitch Control (Spearman, 2018) measures the probability that a team will retain possession of the ball if they pass it to another location on the field. It can evaluate passing options for a player and quantify the probability of success. This methodology for assessing the quality of off-ball positioning provides a framework for examining tracking data to study the game state's short-term evolution.
- **Team Personas.** Team Personas (Meza et al., 2018) is a data-driven unsupervised learning method to identify effective playing styles against certain formations or playing styles. Teams' representation is accomplished applying the Latent Dirichlet Allocation (LDA) algorithm, which assigns a probability to each playing style per team using event data.

To conclude, many different football data types exist that capture tactical, technical, physical, and psychological aspects and meaningful scientific studies leading football clubs to adopt data to serve other purposes. In this concern, data practices have been growing over the years as data availability also increases. Hence, there is an appropriate level of innovation regarding data-centric solutions, pointing to enhancing this field in future years.

“Data is the new oil.”
- Clive Humby

3. DATA

Within the field of football analytics, there are generally two data types, that is, event data and tracking data.

As previously mentioned, a football game is described by a sequence of events on the pitch. In this regard, event data annotate a set of features describing specific discrete events that occur in a game. Supplementarily, on-ball actions are a subset of events captured by these systems that require players to act. Intuitively, a passing event is an action, whereas an event signifying the end of the game is not an action. Typically, event data measures the player acting, body part used, game clock in which the action occurred, start and end location, successfulness, among other features (see Figure 5).

Figure 5 - Event data representation

On the other hand, tracking data records the spatial location of players, referees, and the ball at a high frequency using tracking systems during games, as shown in Figure 6. With such tools, it is possible to perceive the total distance covered, identify the top speed achieved, understand the acceleration direction, study pass availability, measure on-ball pressure, capture tactical structures, et cetera. Additionally, tracking data can be collected using broadcast systems that usually possess a narrow-angle or optical tracking systems that use fixed camera-based technologies in stadiums, therefore capturing a wide-angle.

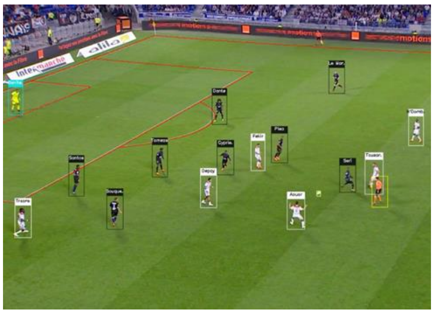


Figure 6 - Tracking data representation

Furthermore, the gathered data that supports football analysis may arise from internal sources, in which case sports entities possess resources to capture data (e.g., fixed cameras and tracking devices) or external sources, in which case data is obtained from outside the organization through specialized providers (e.g., Opta, Wyscout, Stats Perform, SciSports and Second Spectrum).

Data availability decreases as the granularity increases. In concrete terms, due to the high cost of tracking systems, tracking data is available in wealthy leagues or clubs, while event data is more widely and cheaply available. Moreover, tracking data is usually not shared across leagues. However, broadcast tracking systems are filling the gap between usage of event data versus unavailability of tracking data because of its nonexistent in-house implementation costs.

Considering this work's purpose, the developed projects focus exclusively on event data powered by Opta. Nevertheless, event data often lack relevant contextual information about the game, whereas tracking data can strengthen future work.

“Football is a game of mistakes, whoever makes the fewer wins.”
- Johan Cruyff

4. ANALYSIS

In this section, the developed projects within the internship scope will be sequentially presented in-depth.

4.1. PASS SONAR: REPRESENTING PLAYER PASSES BY DIRECTION

4.1.1. Dataset

For the project purpose, it was analyzed event data from Opta, considering exclusively player passes that are not prevenient from kick-offs, freekicks, corners, or goal-kicks throughout the 2019/20 English Premier League season. The dataset consists of 340.287 passes over 380 matches, and the variables used and metadata can be found in Appendix Table 11.

4.1.2. Methodology

In this section, the practical methodology adopted for this data visualization project is exploited. Hence, the implemented procedures for the project development will be subsequently described below.

The data was not incomplete or lacking in certain behaviors in terms of data preparation. Likewise, no outliers were identified facilitating this process. Nevertheless, concerning data inconsistency, there were pass angles beyond 360 degrees. Hence, these values were subtracted by 360.

To achieve the Pass Sonar emerged the need to create a dictionary where keys denote player names while values indicate x and y coordinates of the player's starting position. These coordinates represent the position where the players' polar bar is located in the field. Additionally, substitute players have their coordinate system, i.e., these players are illustrated under the visualization, ordered left-to-right according to minutes played.

Moreover, player passes were aggregated into bins based on the pass angle, where each bin contains a range of 20 degrees, totaling 18 bins representing pass direction. For instance, a pass indicating an angle of 1 or 15 degrees will fall into the same bin since both share the same direction. Afterward, the average pass length and pass frequency were measured in the respective direction or bin per player.

Furthermore, it was necessary to guarantee that players' polar bars are not independent of each other, in other words, the size and color of the bars are uniforms among the players, making the visualization interpretable collectively. In this regard, a function was built that returns the minimum and maximum values of pass angle frequency and pass average distance by selection.

The football pitch was represented using mplsoccer[6] python package, on top of which each players' polar bar was illustrated utilizing matplotlib[7] plotting library.

Subsequently, the python code developed was replicated in Power BI python visual, which, in terms of setup, both software shares the same environment.

[6] https://github.com/davidsousa98/mplsoccer
[7] https://github.com/matplotlib/matplotlib

4.1.3. Results

Through the described methodology, it was possible to produce an interactive visualization capable of identifying the players that make the most and least passes per game, perceiving pass tendencies per player, visualizing the impact of the substitutes, studying passes behavior over time, and analyze unsuccessful passing patterns.

Concerning the data studied, some interesting insights were notable, namely, goalkeepers tend to perform the furthest passes concerning pass distance. On the other side, center forwards usually show the lowest pass frequency while midfielders tend to construct the game, as expected, possessing numerous passes in every direction. Additionally, it is clear to recognize passing patterns by direction, highlighting players that tend to pass forward, backward, or laterally. Therefore, the Pass Sonar can identify the most impactful players throughout each match, concerning the most frequent action in football, that is, passes.

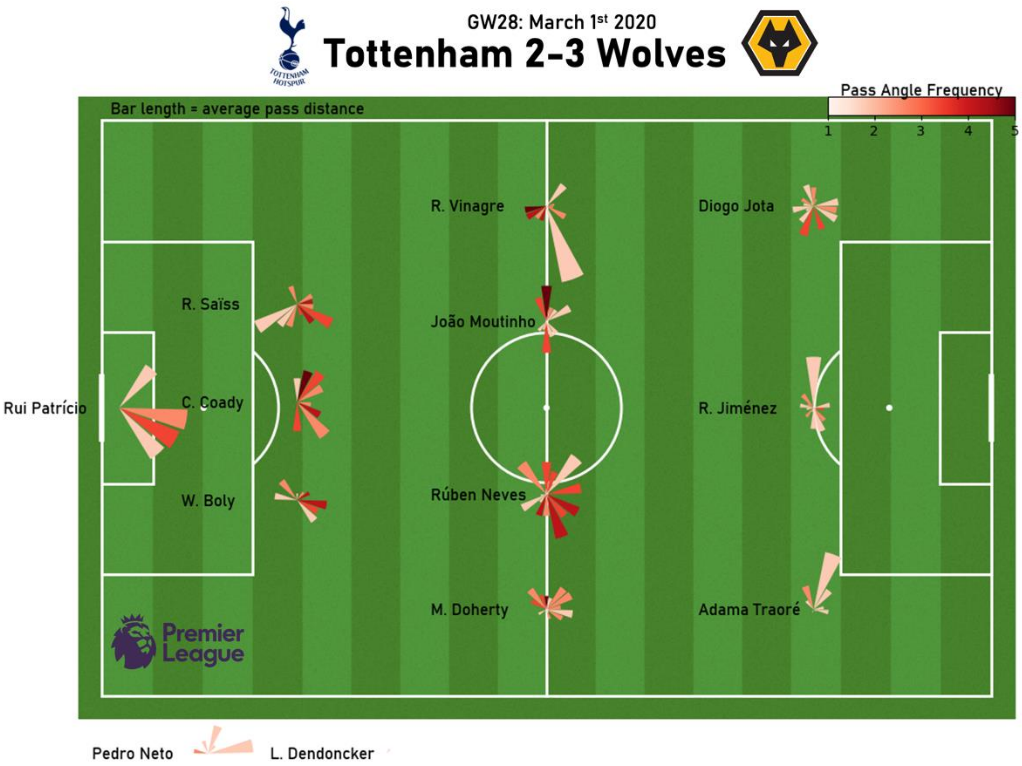


Figure 7 - Pass Sonar: Totthenham vs Wolverhampton

In this particular match during game week 28 of the 2019/20 season (see Figure 7), it is perceptible that, even excluding goal-kicks, Rui Patrício naturally performs the longest passes mainly due to his position in the field. Also, captain Coady does not present any backward pass, as opposed to the full-back Rúben Vinagre who passes mostly backward. Comparing both central midfielders, Rúben Neves was more involved in game creation than João Moutinho, presenting more dispersed passing patterns. Furthermore, there is evidence that Diogo Jota tends to play more centrally than Adama Traoré, since the Portuguese winger also presents passes towards the nearest sideline. In terms of substitute players, Dendoncker, who replaced Jota had no impact in terms of passes, perhaps highlighting the Belgian defensive characteristics, in contrast with Pedro Neto, who replaced the former Barcelona winger.

Additionally, the visualization reacts according to user demands. In other words, the user can filter the visualization by competition, season, team, match, players, position, time played, and complete or incomplete passes. Consequently, as illustrated below, the visualization automatically readapts without additional costs or efforts.

Figure 8 – Representation of Pass Sonar tool

As a result, through this dynamic component, it is pleasant to perceive the behavior of each player in different periods of the game, as a representative case, recognize the most impactful player in the last 10 minutes or, on the other side, identify the player that performed the most amount of unsuccessful passes by time period. Nonetheless, the user can study a particular player or subset of players in-depth either via player name or position, i.e., the visualization can be exploited collectively or individually.

By analyzing Tottenham's visualization, it is noticeable that Spurs had more possession (65%) against Wolves (35%) since, in general terms, Spurs players have a higher passing frequency in all directions. This difference is even aggravated when comparing midfielders from both teams.

Furthermore, it is possible to filter passes per location. In other words, the user can select passes that occurred in the final third of the pitch or during the build-up phase to highlight players that are more impactful in those areas and/or moments of the game. Similarly, through a binary variable that captures if a pass succeeds a backward pass between sectors, it is possible to comprehend how each player reacts to these situations. Also, the user may filter only passes between lines to identify the most creative players and perceive risk-reward passes.

4.1.4. Discussion

The results obtained reveal that it was possible to achieve an effective solution to study player passes and, therefore, provide thoughtful decisions based on facts concerning pre-match analysis.

Moreover, an example of the Pass Sonar's main strengths lies in the fact that once the data which supports the visualization is consistent and updated, the visualization automatically adapts to user demands with no additional costs or efforts. In other words, assuming that the dataset that sustains the visualization consists of the top 5 European Leagues over the past three seasons, Pass Sonar can run between leagues and games efficiently.

Another example of representing player passes in football analytics is through pass networks or pass maps, highlighting connections between players in terms of pass frequencies. From my perspective, these visualizations should be interpreted in a complementary panorama rather than substitutes since both tools have different purposes and, consequently, provide distinct insights.

Furthermore, changing the data granularity from game to season can bring interpretability advantages. Namely, increasing the time window that aggregates several games in a single visualization adds a consistent perspective, perceiving common patterns rather than specific behaviors of a single game.

Notwithstanding, the same methodology can be applied to a different technical aspect of the game in future work. In this way, by replacing the progression method from passes to ball carries, the user can identify decision patterns concerning progressive tendencies by direction (i.e., inside versus outside play), recognize players that progress the furthest with the ball at the feet, as well as identifying ball loss practices by direction. In future work, players and teams that share similar passing trends can be grouped using cluster analysis.

In terms of limitations, the players are represented in a fixed position on the field, which in this case, denotes the starting line-up formation. However, it is common for players to change positions throughout the match. Hence, it would be meaningful to include players' positions when the pass occurred, allowing tactical variations throughout matches.

4.2. VAEP FRAMEWORK: VALUING PLAYER'S ON-BALL ACTIONS

4.2.1. Methodology

The methodology used in this project is the Cross Industry Standard Process for Data Mining (CRISP-DM), one of the most popular knowledge discovery processes. This framework was developed by a consortium of companies involved in data mining (Chapman et al., 2000). The CRISP-DM is considered a de-facto standard process for developing knowledge discovery solutions (Martinez-Plumed et al., 2019). Another data mining framework is SEMMA (Azevedo & Santos, 2008; Shafique & Qaiser, 2014) an acronym for Sample, Explore, Modify, Model, and Assess, developed by the SAS Institute (SAS Institute, 2013). Furthermore, DMAIC is an acronym for Define, Measure, Analyze, Improve, and Control, used in Six Sigma practice (Kubiak & Benbow, 2005). Lastly, the Selection, Preprocessing, Transformation, Data Mining, Interpretation, and Evaluation framework is used in the knowledge discovery in databases process (Fayyad et al., 1996). All these frameworks exhibit common characteristics, and hence, a generic framework closely resembling the CRISP-DM process will be exploited.

The CRISP-DM process was developed in the mid-1990s (Marbán et al., 2009) by a European funded consortium (Martínez-Plumed et al., 2019) composed by DaimlerChrysler, Teradata, OHRA, SPSS and NCR (Azevedo & Santos, 2008; Shafique & Qaiser, 2014; Marbán et al., 2009). The first version of CRISP-DM, providing a uniform framework and guidelines for planning and conducting data mining projects, was published in 1999 (Shafique & Qaiser, 2014).

CRISP-DM builds on previous definitions of knowledge discovery methodologies (Wirth & Hipp, 2000). Aiming to provide a uniform and structured approach to data mining projects development, the CRISP-DM reference model overviews the life cycle of a data mining project, consisting of six well-defined phases, their respective tasks, and outputs (Wirth & Hipp, 2000; Shafique & Qaiser, 2014). However, undoubtedly the field has moved on considerably, with data science being the leading term over data mining. Therefore, CRISP-DM can be readjusted from data mining processes to data science trajectories (Martínez-Plumed et al., 2019). Figure 9 provides a visual overview of the CRISP-DM framework.

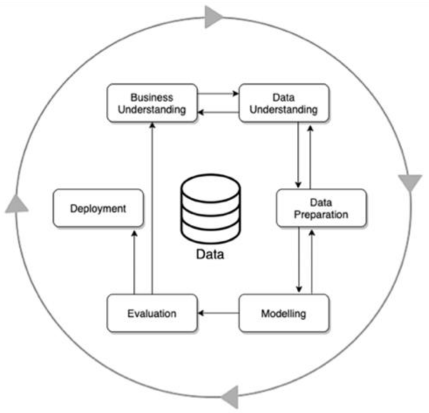


Figure 9 - Most frequent dependencies between CRISP-DM project phases
Adapted from (Chapman et al., 2000)

The sequence of CRISP-DM phases is not strict. In practice, it will often be necessary to backtrack to previous phases and repeat specific tasks as a response to the outcome of each phase (Wirth & Hipp, 2000).

4.2.2. Business Understanding

This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data science problem definition and a preliminary plan designed to achieve the objectives.

Problem Definition

Broadly speaking, most actions in football are performed to increase the chance of scoring a goal, or decreasing the likelihood of conceding a goal. Given that the influence of most actions is temporally limited and goal-dependent hence, one way to assess an action's effect is by calculating how much it alters the chances of both creating and conceding a goal action, that is, a shot possessing expected goal value (xG) above 20% or a goal in the near future. Thus, the effect of an action on creating and conceding was treated separately as these effects may be asymmetric in nature and context-dependent (Decroos et al., 2019).

Valuing an action for a team that possesses the ball requires assessing the change in short-term probability for both creating and conceding as a result of the action a_i moving the game from the state $S_{i-1}$to state S_i. Let $P_{creates}(S_i, x)$ and $P_{concedes}(S_i, x)$ denote the probability of team x respectively creating and conceding a goal action in the near future. The change in creating probability can be computed as:

$$\Delta P_{creates}(a_i, x) = P_{creates}(S_i, x) - P_{creates}(S_{i-1}, x) \tag{1}$$

This change would be positive if the action increased the probability that team x will create a goal action. This change $\Delta P_{creates}(a_i, x)$ was denominated the *offensive value* of an action a_i for team x.

As a visual representation (see Figure 10), Ben Chilwell's long ball clearly changed the state of the game by creating a goal opportunity from near his own box, characterizing a valuable offensive action since it increases Leicester's probability of scoring.

Figure 10 – Video representation of Ben Chilwell's long ball valuable offensive action

Similarly, the change in probability for team x conceding can be computed as:

$$\Delta P_{concedes}(a_i, x) = P_{concedes}(S_i, x) - P_{concedes}(S_{i-1}, x) \tag{2}$$

This change will be positive if the action increases the probability that team x will concede a goal opportunity. Notwithstanding, all actions should aim to decrease the probability of conceding. Therefore, the negation of this change $-\Delta P_{concedes}(a_i, x)$ was denominated the *defensive value* of an action a_i for team x.

As an example (see Figure 11), Sokratis' ball lost inside his defensive box represents a non-valuable action as it increases the likelihood of Arsenal conceding a goal.

Figure 11 - Video representation of Sokratis ball lost

Additionally, Equations 1 and 2 were combined to derive an action's *VAEP value*, that is the sum of action's *offensive value* and *defensive value,* which can be mathematically represented as:

$$\mathrm{V}(a_i, x) = \Delta P_{creates}(a_i, x) + (-\Delta P_{concedes}(a_i, x)) \quad (3)$$

The VAEP framework provides a simple approach to valuing actions independent of the representation used to describe actions. The framework's strength is that it transforms the subjective task of valuing an action into the objective task of predicting the likelihood of a future event naturally (Decroos et al., 2019), allowing to convert everything that happens in the beautiful game to a single currency.

In this way, a valuable action is characterized by increasing the probability of creating a goal opportunity and/or decreasing the probability of conceding, as shown in Figure 12.

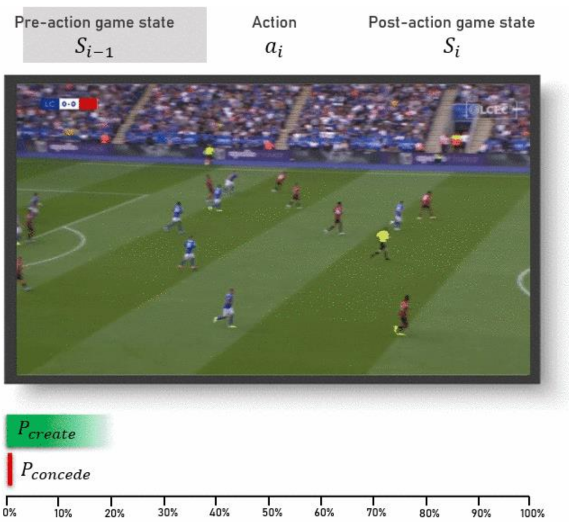


Figure 12 - Video representation of Ben Chilwell's offensive contribution

4.2.3. Data Understanding

During this phase, the data is collected and explored, allowing proper familiarization regarding data quality and interesting subsets or underlying patterns. In sum, the first insights into the available data are formulated.

Data Collection

The collection of data items that support the development of the VAEP framework consists of Opta event data from the English Premier League concerning 2017/18, 2018/19 and 2019/20 seasons. From 1,952,207 events, the dataset was reduced to 1,757,328 on-ball actions, therefore, excluding events such as weather changes, referee decisions, pitch invasion, injured players, ball out, among others. In this manner, each data point represents on-ball actions performed by players, whereas the exploited attributes and metadata can be seen in Appendix Table 12.

Data Exploration

To comprehend the structure of the dataset and the relationship between the predictors and the target variables, a preliminary data analysis was executed. The raw dataset, resulting from the data collection task, comprises 1,140 different games across three seasons, characterized by 51 independent variables, where each game contains $\pm$ 1,540 actions on average. Additionally, two target features are present in the dataset, accounting for the offensive and defensive components of the framework.

In this sense, Opta's frequency of each action type was studied. As illustrated in Figure 13, passes are the most frequent action of the game, representing three-fifths of the dataset. Notwithstanding, the dataset accounts for both offensive and defensive actions.

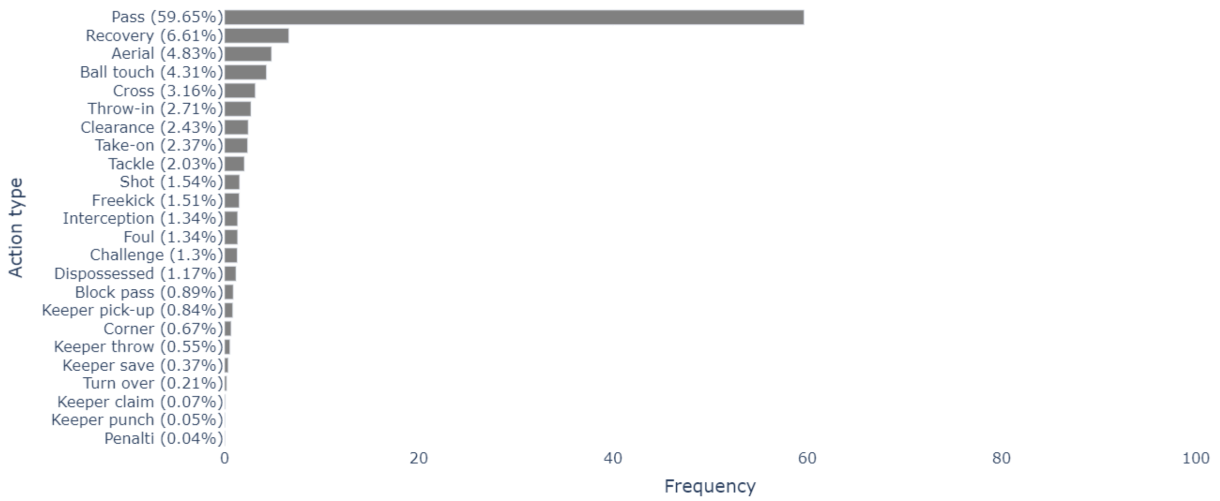


Figure 13 - Frequency of each action type from Opta

Each team's number of offensive and defensive actions was analyzed (see Figure 14). Unsurprisingly, due to Pep Guardiola's possession game, the Citizens perform the highest amount of on-ball actions per game (976), also characterizing the most offensively active team in the English Premier League across 19/20 season. On the other hand, the Saints embody the most defensively active team in the league, averaging 149 defensive actions per game. Additionally, the Magpies represent the team that produces the least on-ball actions per game (644).

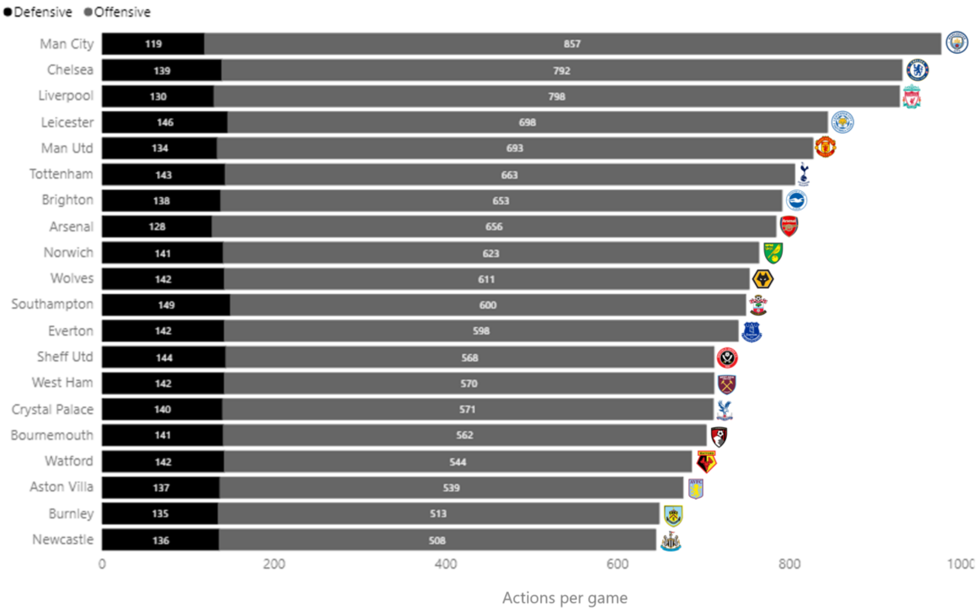


Figure 14 - Offensive and defensive actions per team

Concerning actions' output, goal opportunities are sporadic events in the game. Hence, the dataset is highly imbalanced, with about 2.4% out of over a million actions leading to creating goal opportunities and around 1% conceding, as shown in Figure 15.

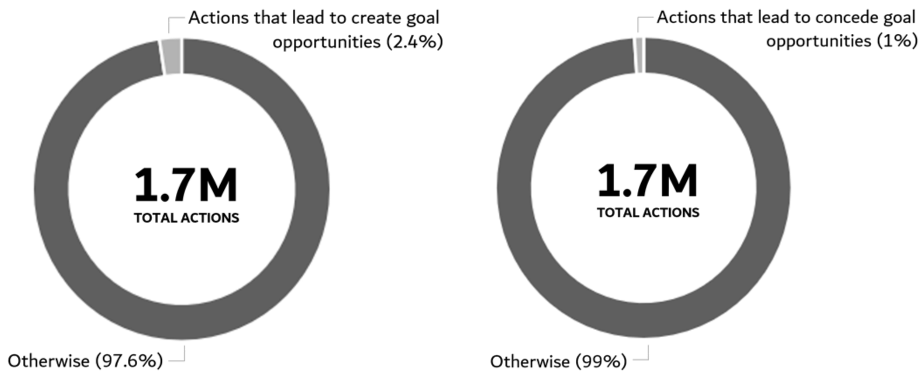


Figure 15 - Distribution of target variables

The following section will tackle this behavior by applying natural and artificial resampling techniques, which either inject data instances from the minority class or remove actions from the majority class.

4.2.4. Data Preparation

This phase encompasses all the steps to construct the processed dataset that will serve as input for modelling the algorithm(s). Usual data preparation tasks include record, feature selection, data cleaning, new attribute construction, and attribute transformation.

Data Cleaning

The dataset that supports the analysis was divided into training set used to fit the model and perform the hyperparameters tuning and testing set applied to access the model's performance and test its generalization capability independently in another set. The training set contains actions from 2017/18 and 2018/19 seasons (66.7%), while the testing set refers to observations concerning 2019/20 season (33.3%). Through the described hold-out method, by fixing the league, there is evidence that the gameplay characteristics are more homogeneous within the same league than several different leagues, without major changes in the distribution of the variables. Additionally, once football evolves over time, the time window has been shortened throughout three consecutive seasons. Also, the distribution of target variables remains consistent across both sets.

Considering that variables are measured in different scales, the model may perceive them differently. Hence, z-score standardization characterizes the number of standard deviations data points above or below the mean value, adjusting values measured on different scales to a common scale (i.e., standard deviations). A Standard Scaler was also fitted to the training set, which can standardize the dataset by removing the mean μ and scaling each feature x to unit variance σ, per Equation 4.

$$Z_x = \frac{x_i - \mu_x}{\sigma_x} \tag{4}$$

In this manner, Standard Scaler can scale and center each feature independently. This task was undertaken since algorithms, such as Logistic Regression, require the data to be standardized to avoid unit scaling effects. Data standardization was done by applying Python's *sklearn.preprocessing.StandardScaler* class. After finding the mean and standard deviation of the training set, both training and testing sets were transformed.

In terms of incoherent values, actions outside the Opta (x, y) coordinate system by a substantial difference (see Figure 50 of the Appendix) range from 0 to 100, were excluded from the study.

This phase also includes the identification of outliers, which are values against the pattern of the data, in some cases, indicating suspicions that a different mechanism originated these observations. In this way, having searched for anomalous values, a univariate outlier analysis was undertaken by exploiting boxplots and histograms of the predictive features to identify extreme values that deviate from the remaining data points. There is an indication that, when analyzing predictors individually, there are no instances that lie an abnormal distance from some other cases. Thus, no outliers were identified.

Concerning missing values, observations in which the player acting was unknown were removed from the analysis once several features were lacking. Additionally, the data provider only supplies angles across passes. Hence, action angles beyond passes have been replaced by zero.

Feature Engineering

Predictive algorithms benefit from Feature Engineering since not every feature is relevant to predicting the target. In this thesis, Feature Selection methods were employed, which aim to select features according to their discriminatory ability (Khalid et al., 2014), namely, Spearman Correlation Analysis, Recursive Feature Elimination, and Ridge Regression were considered for improving models' performance. Additionally, Feature Extraction was also implemented, which transforms the original features to generate derived features that are relevant and non-redundant (Khalid et al., 2014), facilitating the subsequent processes.

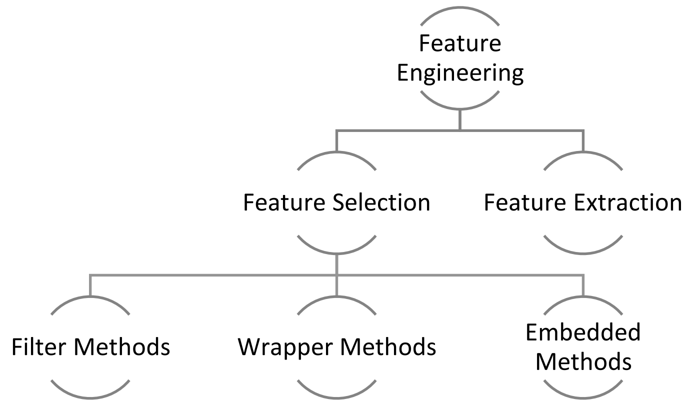


Figure 16 – Hierarchical Structure of Feature Engineering

Feature Selection methods can be classified into filter, wrapper, and embedded methods (Khalid et al., 2014), as illustrated in the figure above. Spearman correlation was analyzed in terms of filter methods, which does not incorporate specific machine learning algorithms in contrast to the remaining methods mentioned. This correlation differs from Pearson's since it evaluates non-parametric correlations through ordinal information and score of pairs, contrary to Pearson's, which evaluates parametric correlations. Through this method, two highly correlated variables suggest redundancy since there is an indication that both may contain the same information. Nevertheless, the pair of variables that possess the highest correlation are the start and end locations of action (≈85%). However, through previous research and context knowledge, there is evidence that these are relevant features to explain the likelihood of creating and conceding goal actions once two actions with the same distance covered might have different impacts on the game depending on its location.

Recursive Feature Elimination (RFE) was implemented using *sklearn.feature_selection.RFE* class as a wrapper feature selection method, RFE evaluates on a specific machine learning algorithm to find the optimal number of features. Concretely, given an external estimator that assigns weights to features, the goal of RFE is to select features by recursively considering smaller and smaller sets of features. Firstly, the estimator is trained on the initial set of features, and the importance of each feature is obtained. Then, the least important features are pruned from the current set of features. That procedure is recursively repeated on the pruned set until the desired number of features is eventually reached. In this step, logistic regression was the external estimator applied to consider its low complexity and running time.

Concerning embedded methods, which contain their built-in feature selection processes, Ridge Regression was performed using *sklearn.linear_model.RidgeClassifier* that converts the target values into {-1, 1} and then treats the problem as a regression task applying L2 regularization. As a result, the method returns the coefficients associated with each feature in the decision function.

Concerning constructing features, from the original dataset arose the need to obtain the action end location on a case-by-case basis according to the action type. Subsequently, the distance covered was obtained once known the start and end location of an action. Additionally, in terms of complex features, the distance and angle towards the opponent's goal were measured concerning both the start and end location of an action. Moreover, concerning shots off-target, the distance towards the center of the goal was quantified. Since teams often adapt their playing style to the current scoreline (Decroos et al., 2019), the goal difference before an action is performed was denoted.

In order to predict the short future, that is, the likelihood of both scoring and conceding a goal in the following k actions, through previous research and evaluating model performance, there is evidence that the short past is relevant to predict the future. Said that, a game state S_i is a group of consecutive actions. For action a_i instead of defining features based on the entire current game state $S_i = [a_1, a_2, \dots, a_i]$ it was only considered the previous three actions $[a_{i-2}, a_{i-1}, a_i]$.

Previous approaches focus exclusively on the goal while creating the target feature. However, goals are a very rare occasion of the game. On average, less than 1% of the actions result in a goal. Also, goals are directly dependent on the finishing capacity of the attacker, i.e., teams may be able to unlock the opponent's defense and create a clear goal opportunity. Still, the finishing ability of the attacker is unable to materialize it. Hence, to mitigate these obstacles and give credit to actions, expected goal models (xG) which assign a probability of each shot to result in a goal were exploited, changing the predictive objective to creating or conceding goal opportunities rather than exclusively focusing goals, therefore making the prediction broader.

Hence, the predictive task simplifies into two separate binary probabilistic classification problems with identical inputs but different labels. For the first classification problem of estimating $P_{creates}(S_i, x_i)$, it was assigned a game state S_i a positive label (=1) if the team possessing the ball after action a_i created a goal opportunity in the subsequent k actions, that is, a shot possessing an expected goal value (xG) above 20% or a goal, and a negative label (=0) otherwise. Similarly, for the second classification problem of estimating $P_{concedes}(S_i, x_i)$, it was assigned a game state S_i a positive label (=1) if the team possessing the ball after action a_i conceded a goal opportunity in the subsequent k actions and a negative label (=0) in all other cases.

In both binary classification problems, k is a user-defined parameter that represents how far ahead in the future we observe to determine the effect of an action. In this thesis, based on domain knowledge and preliminary experiments, k=10 was considered.

In addition, the evaluation of each feature set was performed using a logistic regression algorithm due to its simplicity and fast running time. The results can be accessed in Table 1.

Feature set	Create target Brier	Create target AUC	Concede target Brier	Concede target AUC
All features	0.02017	0.7758	0.00950	0.7558
Ridge features	0.02018	0.7759	0.00950	0.7558
RFE features	0.02019	0.7760	0.00950	0.7542
No game context features	0.02018	0.7753	0.00950	0.7512
No game state features	0.02041	0.7734	0.00979	0.7518
Baseline	0.02427	0.5	0.01025	0.5

Table 1 - Feature selection results

As shown, several feature sets were compared against the baseline, which always predicts the mean class probability, i.e., if the dataset contains 1.5% positive examples, the model always predicts 0.015. Even though the results reflect a marginal difference, there is evidence that retaining all features is the most effective representation, supported by the a priori knowledge acquired from previous studies.

Resampling techniques

As previously mentioned, during the definition of the target variable, goals are a very rare occasion of the game, on average, less than 1% of the actions result in short-term goals. Consequently, the dataset is highly imbalanced. Many examples represent one class while only a few describe the other. Hence, through previous studies, there is evidence that imbalanced learning was shown to hurt the performance of standard classifiers (Japkowicz & Stephen, 2002). Several resampling techniques described below were exploited to mitigate the imbalanced class problem (see Table 2).

Resampling Technique	Description	Python Package
Natural Oversampling	Consists of injecting new data from other leagues which belong to the minority class. According to *UEFA Association Coefficients*, Ligue 1 (France) and Jupiler Pro League (Belgium) were tested once the goal was to append homogeneous leagues.	N/A
Natural Undersampling	Involves excluding observations from the majority class, which occurred in the defensive half of the pitch when predicting $P_{creates}(S_i, x_i)$ once, in general terms, actions that end closer to the opponent's goal are more likely to result in goal opportunities.	N/A
Hybrid Method	Comprises natural oversampling and undersampling jointly.	N/A
Random Oversampling	Generates exact replicas of the minority class, i.e., create clones of actions based on the existing ones.	*imblearn. over_sampling. RandomOverSampler*
Random Undersampling	Randomly excludes observations from the majority class.	*imblearn. under_sampling. RandomUnderSampler*

Synthetic Minority Oversampling Technique (SMOTE)	Using SMOTE, a subset of data is taken from the minority class as an example, and then new synthetic similar instances are created.	*imblearn. over_sampling. SMOTE*
Adaptive Synthetic Oversampling (ADASYN)	ADASYN is a similar approach compared to SMOTE but with a minor improvement, that is, after creating the similar synthetic instances, a small random value is added. In other words, instead of the sample being linearly correlated to the parent, they possess an additional variance component.	*imblearn. over_sampling. ADASYN*
Cluster-Based Undersampling (K-means)	Keeps K majority samples by fitting K-means algorithm with K clusters to the majority class and using the coordinates of the K cluster medoids as the new majority samples.	*imblearn. under_sampling. ClusterCentroids*
K-Nearest Neighbors Undersampling (KNN)	Removes instances of the majority class if the count of its neighbors is greater or equal to the threshold (Beckmann et al., 2015).	*github.com/ marcelobeckmann/ knnund*

Table 2 - Description of implemented resampling techniques

From the results obtained (see Table 3), there was no improvement compared to prior Neural Networks results without any transformation. Typically, after implementing previously declared resampling techniques, the model enhanced predictive capability respecting the minority class. However, on the other hand, the model tends to increase the predictive error while predicting majority class instances. Thus, the tradeoff was not carried forward.

Resampling technique	**Create target**	
	Brier	**AUC**
No technique applied	0.01989	0.7771
Natural undersampling	0.01988	0.7731
KNN undersampling	0.02012	0.7764
Hybrid method	0.02046	0.7751
Natural oversampling (Jupiler League)	0.02054	0.7768
Natural oversampling (Ligue 1)	0.02104	0.7779
Natural oversampling (Jupiler + Ligue 1)	0.02295	0.7784
SMOTE	0.02438	0.7489
Random oversampling	0.02440	0.7774
ADASYN	0.02477	0.7434
Random undersampling	0.09043	0.7735

Table 3 - Resampling techniques results

4.2.5. Modelling

In this phase, the appropriate data mining task (e.g., classification, clustering, regression) is identified, and corresponding data mining and machine learning algorithms are selected, implemented, and fine-tuned. Several techniques are considered for modelling the same problem, which will be evaluated to obtain the best-performing model.

The training set is used to retrieve discoveries, make decisions, tune the hyperparameters, and train the classifier during the modeling stage. On the other hand, the testing set enables the assessment of the final classifier performance on unseen data. It reduces the bias of developing an overfitted model in which results are too adjusted to a particular data set. The classifier is likely to fail when predicting unseen observations (Leinweber, 2007). Naturally, the modelling phase is executed in pair with the evaluation phase, representing a back-and-forth process, as shown in Figure 17.

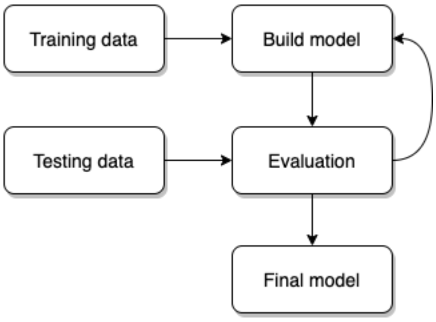


Figure 17 - Modelling steps
Adapted from (Chapman et al., 2000)

The classifiers used during this process and their respective specifications are described according to Table 4.

Algorithm	**Python Package**	**Version**
Logistic Regression	*sklearn.linear_model.LogisticRegression*	0.23.2
Neural Network	*sklearn.neural_network.MLPClassifier*	0.23.2
Random Forest	sklearn.ensemble.RandomForestClassifier	0.23.2
XGBoost	*xgboost.XGBClassifier*	1.4.2

Table 4 - Model specifications

Logistic Regression

In logistic regression, the goal is to test a predictor or set of predictors of a categorical dependent variable. Independent variables can be nominal, ordinal (ranked), interval, or ratio level (or continuous) data. This binary problem can be extended to multinomial regression by using a set of predictors on a set of dependent variables.

The S-shaped curve *logistic function*, Equation 5, describes the mathematical form on which the *logistic model* is based (Kleinbaum et al., 2010), as can be seen below.

$$f(z) = \frac{1}{1 + e^{-z}} \tag{5}$$

The logistic function, *f(z)*, ranges between 0 and 1 (see Figure 18) is the primary reason the logistic model is so popular. The model is designed to describe a probability, always some number between 0 and 1 (Kleinbaum et al., 2010).

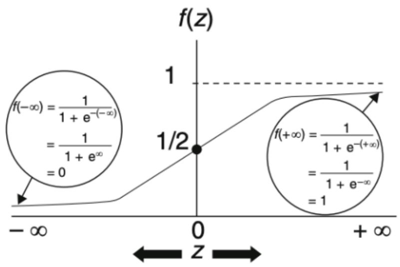


Figure 18 - Logistic function ($0 \leq f(z) \leq 1$)
Adapted from (Kleinbaum et al., 2010)

To obtain the logistic model, for k equal to the number of independent variables and i ranging from 1 to k, z represents the linear sum between intercept value α plus the regressors X_i times their effects β_i whereas constant term α and coefficients β_i denote unknown parameters.

$$z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \implies f(z) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}} \tag{6}$$

In this way, the parametric logistic model uses maximum-likelihood estimation to retrieve the estimators for the unknown parameters over the training samples. In football terms, focusing on the offensive component of VAEP, such probability gives the likelihood of creating a goal action for the collection of variables X_1 through X_k. In practice, the dependent variable Y is dichotomous, taking on one (i.e., the positive outcome, or success) and zero (i.e., the negative outcome, or failure). Hence, the probability for actions to result in goal opportunities can be seen as:

$$P(Y = 1 |\, X_1, X_2, \dots, X_k) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}} \tag{7}$$

Based on sigmoid function, these models benefit by not requiring a high computational effort and, as a White Box model, its results are highly interpretable. On the other hand, the results of its linear boundaries can be easily surpassed by more complex algorithms.

Feed-Forward Neural Network

Artificial Neural Network (ANN) model involves computations and mathematics, which simulate the human-brain processes (Malekian & Chitsaz, 2021). ANNs are composed of processing units called neurons, or nodes, and weighted connections between those neurons, analogously to the human brain's synapses. Artificial neurons are organized into input, output, and, in most cases, also hidden layers. Input layer's neurons are called input neurons and, similarly, output and hidden layers' neurons are respectively called output and hidden neurons (Brown & Mues, 2012).

Several different ANN architectures are present in the literature. The most widely used is the Feed-Forward Multilayer Perceptron (Brown & Mues, 2012). In Feed-Forward Neural Networks (FNN), each layer is fully connected, meaning that every neuron in a layer connects to all nodes in the previous layer. In other words, all possible intra-level connections between two adjacent layers are established, as shown in Figure 19.

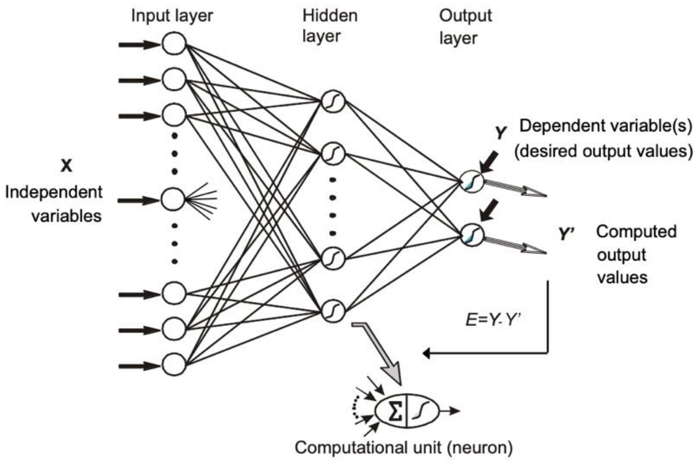


Figure 19 - Schematic of three-layered feed-forward neural network
Adapted from (Park & Lek, 2016)

The input units pass the assigned values of the independent variables to the first hidden layer, serving as inputs to the network. On the other hand, the hidden units act as feature detectors (Tu, 1996). The procedure of inputs' processing and output generation at each hidden neuron is summarized in Figure 20. In sum, this process multiplies the input values with its weight vector, sums the multiplied values, applies its activation function (e.g., sigmoid function) to the weighted sum, and returns the value computed by the output function (e.g., identity function) to the next layer. There can be any number of hidden units, while there should be one input unit for each input variable. Finally, the output units represent the predicted outputs of the network.

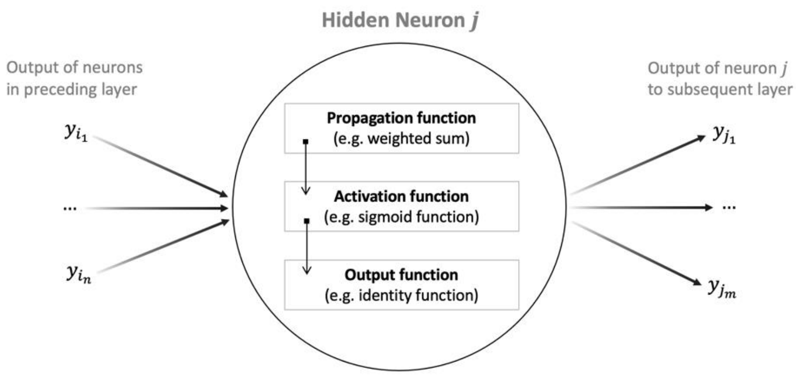


Figure 20 - Input processing and output generation in a hidden neuron
Adapted from (Kriesel, 2007)

FNN are trained to perform input-to-output non-linear mappings through the use of an error correction procedure, that is, backpropagation of error (Rumelhart et al., 1986). The backpropagation algorithm is summarised below. Nonetheless, implementation details can be found in most neural network books (e.g., Bishop, 1995).

1. initialise network weights (e.g., random initialization);
2. present first input vector from training data to the network;
3. propagate the input vector through the network to obtain an output;
4. weights are iteratively adjusted to minimize an objective function (i.e., loss function), for instance, the Mean Square Error;
5. propagate error signal back through the network;
6. adjust weights to minimise overall error per Gradient Descent or other optimization methods, namely RMSprop and Adam;
7. repeat steps 2-7 with the following input vector until the overall error is satisfactorily small.

As stated by the Universal Approximation Theorem, assuming the nonlinearity of the activation function for the hidden neurons, FNNs can arbitrarily approximate any nonlinear relationship between the dependent and independent variables (Tu, 1996). In other words, the decision boundaries for FNNs can be nonlinear, granting these models more flexibility when compared to order approaches, such as Logistic Regression (Dreiseitl & Ohno-Machado, 2002).

Random Forest

As the name suggests, Random Forests is a nonparametric tree-based algorithm that elevates the concept of Decision Trees, as a combination of tree predictors that operate as an ensemble (Breiman, 2001). In general, ensemble methods combine the predictions of some base learners to improve generalization ability and robustness compared to a single base learner's performance.

Two popular ensemble methods are (Sutton, 2005): (1) boosting methods, where weak base learners are iteratively built from weighted training samples. In each iteration, the weights are adjusted to give increasing weight to cases that were misclassified in the previous iteration; (2) and bootstrap aggregation, or bagging methods, where some base learners are independently trained on bootstrap samples drawn from the available data, and then their individual predictions are aggregated to obtain the overall prediction.

For the particular case of Random Forests, each independent tree base learner is grown to full size (i.e., no pruning) and fitted on a training sample, drawn with replacement, that is the same size as the original training dataset (i.e., bootstrap sample). Additionally, when constructing the tree base learners, the best split for each node is found based on a random subset of all possible input features (Brown & Mues, 2012). By introducing these two sources of randomness, the goal is to improve the model stability and reduce both the overfitting tendency and the prediction variance compared to single Decision Trees (Sutton, 2005; Buskirk, 2018). In this manner, several trees work independently where each tree is differentiated by receiving distinct training samples and/or constructed under a random subset of features. Subsequently, the results of individual trees are combined to obtain the final prediction (see Figure 21).

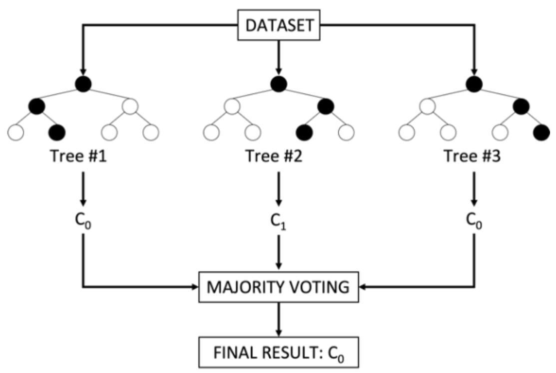


Figure 21 - Random forest representation

Due to its complexity, as a Black Box model, Random Forests suffer from a lack of interpretability and few control on the model's operability. Also, these models require more training time and computational effort when compared to Decision Trees and Logistic Regression. On the other hand, Random Forests can solve both classification and regression problems. These models do not require predictive features to be previously scaled and are robust to noisy data (e.g., missing values and outliers). Furthermore, they are very stable and powerful algorithms that tend to provide good results in a wide range of problems since they can handle high dimensional datasets well and benefit from ensemble and bagging techniques that reduce overfitting and improve accuracy.

XGBoost

The final classifier discussed in this report is called Extreme Gradient Boosting, or as it is commonly denoted, XGBoost.

Similar to Random Forest, XGBoost is also a decision-tree-based ensemble algorithm, a highly effective and widely used machine learning method (Chen & Guestrin, 2016). However, this supervised algorithm performs well in overall problems once it derives from the boosting concept, which aims to create a strong classifier based on weaker classifiers. By adding models on top of each other, the errors of the previous model are corrected by the following predictor (Schapire, 2003). Then, as the name suggests, the gradient aspect of the algorithm looks to minimize the residual loss of adding the last prediction by fitting new residuals to the appended model instead of just assigning different weights in each iteration. In this manner, for a given dataset, the tree ensemble model uses n addictive functions to predict the output, where f_k corresponds to an independent tree (shown in Figure 22).

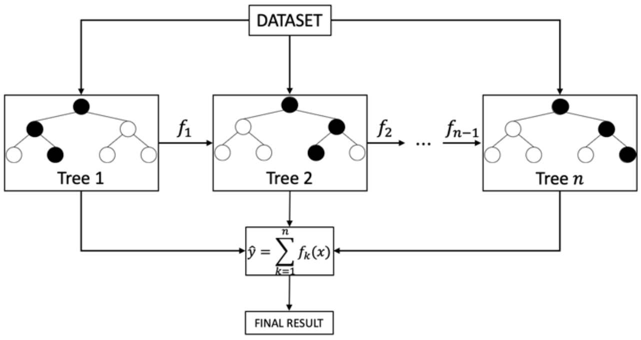


Figure 22 - Extreme Gradient Boosting representation

Ultimately, the impact of XGBoost, designed to be highly efficient, flexible, and scalable, has been widely recognized in several machine learning and data mining challenges. Previous studies demonstrate that XGBoost gives state-of-the-art results on many problems (Chen & Guestrin, 2016). Furthermore, this tree boosting algorithm provides low running time, allows for the parallelization of trees, stands efficient when handling noisy data, and uses regularization that reduces the risk of overfitting. Notwithstanding, since XGBoost classifiers are Black Box models, the prediction process interpretability is highly difficult.

Building and Tuning Models

Hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. For decades, the de-facto standard for hyperparameter optimization in machine learning has been grid search, which is simply an exhaustive searching through a manually specified subset of the hyperparameter space of a learning algorithm (Ghawi & Pfeffer, 2019).

To identify the best combination of parameters used in each model, hyperparameter tuning was performed by applying Python's *sklearn.model_selection.GridSearchCV* class to the training set, from which *sklearn.model_selection.StratifiedKFold* was implemented.

During this phase, cross-validation was applied as a model validation technique for assessing how the model will generalize to an independent dataset. A common type of cross-validation is k-fold cross-validation, where the original sample is randomly partitioned into k equal-sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is repeated k times, with each of the k subsamples used exactly once as the validation data. Hence, the k results can be averaged to produce a single estimation (Ghawi & Pfeffer, 2019).

	Fold 1	Fold 2	Fold 3	...	Fold K
Split 1	Test	Train	Train	...	Train
Split 2	Train	Test	Train	...	Train
Split 3	Train	Train	Test	...	Train
...	...	...	...	...	...
Split K	Train	Train	Train	...	Test

$$\text{Classifier Performance} = \frac{1}{K}\sum_{i=1}^{K} \text{Performance}_i$$

Figure 23 - K-Cross validation with average classifier performance

Considering that target variables are highly imbalanced, stratified 5-fold cross-validation was applied to guarantee that both classes are present in each fold during the hyperparameter tuning process, maintaining representativeness and increasing the robustness of results. In addition, through the implementation of 5-fold cross-validation arises the advantage of evaluating the performance of each fold collectively based on multiple validation sets.

By default, parameter search uses the score function of the estimator to evaluate a parameter setting. The negative brier score was implemented because all scorer objects follow the convention that higher return values are better than lower return values. It measures the mean squared difference between the predicted probability and the actual outcome. Therefore, as an error metric, the lower, the better.

The results of this phase with the respective optimal values for each model implemented can be accessed in Table 13 of the appendix. Notwithstanding, running a grid search in random forests naturally remit to full growth trees, leading to overfitting the training data and losing its generalization capability to unseen data. To avoid this behavior, (1) the number of trees in the forest, (2) the maximum depth of the tree, and (3) the number of features to consider when searching for the best split were tested apart by identifying the cutoff moment in which the ROC AUC score of each forest representation stops increasing across the testing set.

4.2.6. Evaluation

After implementation and parameter calibration, the algorithms' performance must be thoroughly evaluated. Furthermore, the constructed architecture should be reviewed during this phase to guarantee that the project's objectives, defined during the business understanding phase, are being achieved.

Model Assessment

To access the model's performance, it was analysed the brier score, roc auc score, calibration plot, classification report and confusion matrix jointly.

The brier score was applied using *sklearn.metrics.brier_score_loss* to measure the accuracy of the probabilistic predictions. This function returns a score of the mean square difference between the actual outcome and the predicted probability of the possible outcome. Hence, the brier score loss ranges between zero to one and the lower values indicate more accurate predictions.

Additionally, the Area Under the Receiver Operating Characteristic Curve (AUC) was utilized by applying *sklearn.metrics.roc_auc_score* class that evaluates how well the classifiers can distinguish positive examples from negative examples. The AUC score varies from 0 to 1. When AUC is approximately 0, the model reciprocates the classes. In other words, the model predicts negative class as a positive class and vice versa. When AUC is approximately 0.5, the model has no discrimination capacity to distinguish between classes. Finally, AUC of 1 represents an ideal measure of separability, meaning that the model perfectly distinguishes between positive and negative instances.

Since the brier score is an error measure, lower values suggest that the prediction is closer to the ground truth, representing the most desirable scenario. In contrast, respecting the AUC score, higher values indicate that the model is more capable of separating between both classes. In this way, according to Table 5, the XGBoost model outperforms the remaining models.

	Create target		**Concede target**	
Model	**Brier**	**AUC**	**Brier**	**AUC**
XGBoost	0.0197	0.781	0.0094	0.764
Neural Network	0.0199	0.777	0.0095	0.758
Random Forest	0.0199	0.770	0.0095	0.744
Logistic Regression	0.0202	0.777	0.0095	0.759
Baseline	0.0243	0.5	0.0103	0.5

Table 5 - Modelling results (Brier and AUC scores)

Moreover, the calibration plot was employed to compare how well the probabilistic predictions of different classifiers are calibrated, characterizing the confidence on the prediction over the distribution of the target variable. Therefore, from Figure 24, which compares the fraction of positives with the mean predicted values, there is evidence that models overestimate the probability of creating goal actions from the game states representation since generally, the fraction of positives is above the mean predicted values. In this sense, models closer to the perfectly calibrated line indicate better calibrations. Thus, Neural Networks and XGBoost are the two most calibrated models.

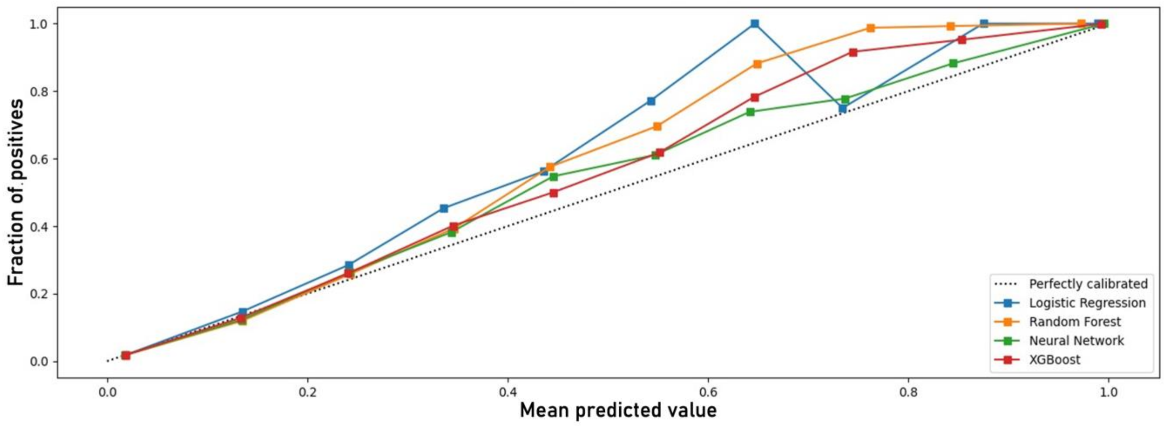


Figure 24 - Calibration curves of the studied models

To complement, the classification report, which presents a text report showing the main classification metrics, such as precision, recall, f1-score, support, was exploited using *sklearn.metrics.classification_report*. Also, the confusion matrix that compares the predicted values with the ground truth across a final frequency table supported models' evaluation process by employing *sklearn.metrics.confusion_matrix*. As shown in the Appendix, the predicted values are represented in columns, while rows illustrate the real values. The elements in the diagonal denote the correctly predicted observations.

Furthermore, the presented metrics were compared across training and testing sets to guarantee that models do not overfit the training data and lose generalization to unseen data. In this way, discrepant scores in both sets indicate that the model is likely to suffer overfitting.

Finally, the Gradient Boosting algorithm outperforms the remaining models by analyzing the described metrics collectively. Hence, XGBoost results were conducted in this report.

4.2.7. Deployment

This phase focuses on organizing, reporting, and presenting the discovered knowledge so the interested parties can use it. Moreover, this phase can also entail the integration of the proposed architecture into another system and subjacent monitoring and maintenance.

<u>Player Analysis</u>

The described methodology assigns a value to each action therefore, it is possible to quantify sequences of play during games. As an illustrative example, Figures 25 and 26 represent a corner from Bournemouth that effectively resulted in goal. As can be seen, the framework gives high importance to Fraser's assist, which directly creates the goal opportunity, and Wilson's shot materializes the goal opportunity. In the opposite direction, King's lateral pass has a negative value which can be justified since the pass end location is further away from the opponent's goal than the starting position. In this sense, the most and least valuable actions during a 90-minute match are noticeable by the player or team, which can posteriorly be utilized to identify video plays.

Figure 25 - Video representation of Bournemouth's goal

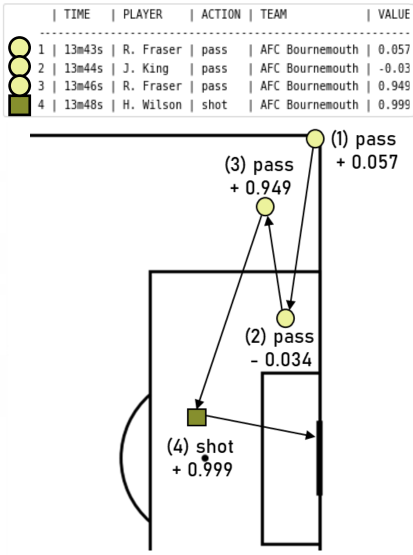


Figure 26 - Valuing on-ball actions

In addition, it is possible to aggregate individual action values into player ratings for multiple time granularities and several dimensions, which was accomplished according to Decroos et al.'s (2019) definition. Regardless of the time frame, the rating is obtained similarly. Furthermore, since spending more time on the pitch offers more opportunities to contribute and perform more on-ball actions, the player's rating was computed per 90 minutes of game time to enhance equity between players. Given a time frame T and player p, the player's rating was calculated as:

$$\text{rating(p)} = \frac{90}{m} \sum_{a_i \in A_p^T} V(a_i) \tag{8}$$

where A_p^T is the set of actions player p performed during time frame T, $V(a_i)$ is computed according to Equation 3, and m is the number of minutes played during T. This player rating captures the average net goal difference contributed to the player's team per 90 minutes.

Once valued every on-ball action, aggregating football players' action values becomes noticeable player's contribution to their team and establishes comparisons of player performance levels throughout the whole season. According to the framework, figure 27 illustrates the highest valued starting XI, mainly characterized by Man City and Liverpool players, runners-up and champions, respectively.

Figure 27 - Premier League 2019/20 highest rated XI

Notwithstanding, players have different usage rates throughout the whole season. Consequently, it wouldn't be adequate to compare a player such as Trent (38 appearances) with Fosu-Mensah (3 appearances) even if both shared similar ratings per 90 minutes. Said that two-dimensional (2D) K-means clustering was performed considering minutes and games played by the player. Through hierarchical clustering dendrogram using the Ward linkage method, there was evidence that players were aggregated into three clusters, high, medium, and low usage rates. Finally, players belonging to the low usage rate cluster were excluded from the analysis.

Through Figure 27, the Citizens Kevin de Bruyne and Riyad Mahrez are the standout players to perform valuable on-ball actions. In addition, the full-back Trent Alexander-Arnold can be as impactful as players in more attacking positions, such as Bruno Fernandes or Sergio Aguero. Unsurprisingly, central backs tend to be less impactful even though the framework contains both offensive and defensive components. Furthermore, the framework also values goalkeeper's on-ball actions, which various metrics (e.g., Goal Contributions and Expected Goals) do not consider once mainly emphasize goals, assists or shots.

By aggregating players per sector, the framework can highlight the uppermost performing players who typically share the same role (see Figure 28).

ATTACKERS SEASON 19/20

	PLAYER	TEAM	RATING
1	R. Mahrez	Man City	1.14
2	S. Agüero	Man City	0.85
3	Gabriel Jesus	Man City	0.76
4	T. Abraham	Chelsea	0.75
5	Mo Salah	Liverpool	0.74

MIDFIELDERS SEASON 19/20

	PLAYER	TEAM	RATING
1	K. de Bruyne	Man City	1.23
2	B. Fernandes	Man United	0.83
3	David Silva	Man City	0.78
4	C. Hourihane	Aston Villa	0.74
5	R. Snodgrass	West Ham	0.73

DEFENDERS SEASON 19/20

	PLAYER	TEAM	RATING
1	T. A-Arnold	Liverpool	0.82
2	A. Robertson	Liverpool	0.58
3	A. Mohamady	Aston Villa	0.48
4	V. van Dijk	Liverpool	0.47
5	Marcos Alonso	Chelsea	0.46

GOALKEEPERS SEASON 19/20

	PLAYER	TEAM	RATING
1	M. Dúbravka	Newcastle	0.68
2	T. Heaton	Aston Villa	0.67
3	M. Ryan	Brighton	0.61
4	B. Leno	Arsenal	0.58
5	A. Ramsdale	Bournmouth	0.58

Figure 28 - Top 5 performing players per sector

Concerning attackers, the blues from Manchester hold the spotlight as the Citizens feature the top 3 attackers during 2019/20 runners-up season, i.e., Riyad Mahrez (top 2 big chances created – 18)[8], Sergio Aguero (top one minutes per goal – 90.9) and Gabriel Jesus (top two shots on target – 48).

In terms of midfielders, the three most impactful players belong to Manchester rivals, Kevin de Bruyne (top 1 goal contributions - 13 goals and 20 assists), Bruno Fernandes (top 1 xG per 90 minutes out of midfielders – 0.47), and David Silva (top 10 chances created per 90 minutes – 2.41), each having characteristics to unlock any game by creating goal opportunities or even materializing them.

The champions, Liverpool, emerge through the defensive sector, Trent Alexander-Arnold (top 1 successful crosses - 81), Andrew Robertson (top 3 attacking vertical passes – 392), and Virgil van Dijk (top 1 accurate passes – 2903) represent three out of the top five defenders.

Concerning goalkeepers, usually, this position is represented by teams that tend to frequently concede goal opportunities, consequently possessing a more active goalkeeper in the game. In this sense, Martin Dúbravka (top 1 saves – 140) is the highest-rated player.

[8] Statistics compared to all players in the league with more than 100 minutes played.

Furthermore, it is relevant to perceive league tendencies per position by aggregating players in the most frequent position used throughout the whole season, Figure 29, in which case the highest-rated player for each position is highlighted.

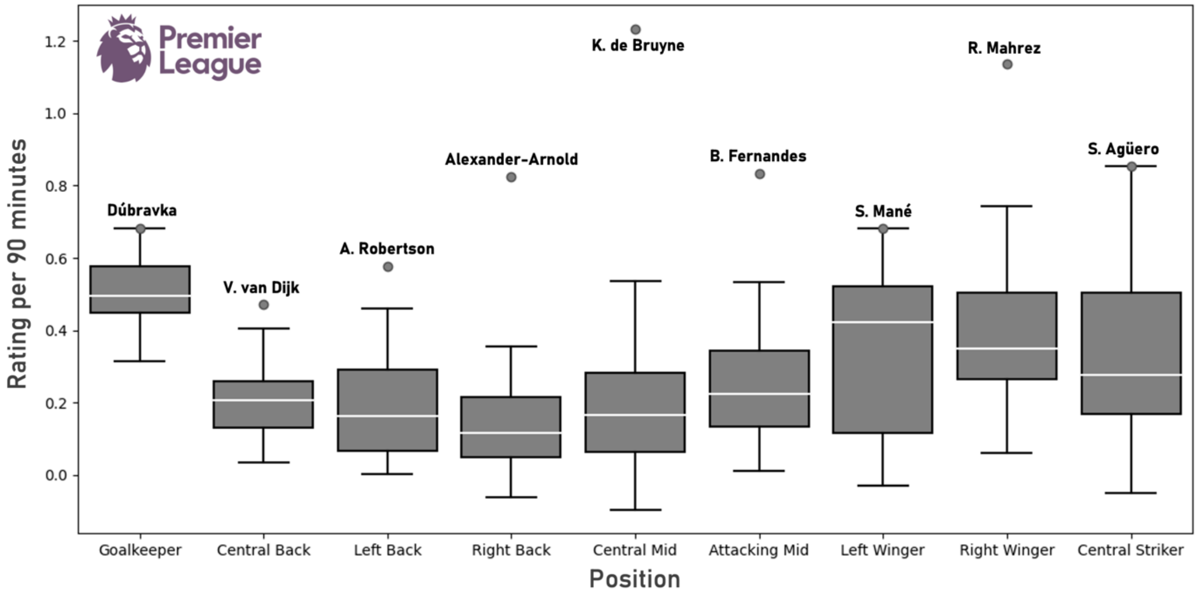


Figure 29 - VAEP rating per position

Expectedly, more advanced positions on the pitch are typically associated with higher ratings besides goalkeepers. This behavior can be explained once goalkeepers are not penalized for conceding goals when they do not perform any on-ball action, even if they are directly involved in the goal conceded due to poor positioning or inappropriate off-ball approach. Additionally, from an offensive perspective, due to the goalkeeper's positioning on the pitch and as seen in the Pass Sonar Project, goalkeepers tendentiously perform the furthest passes therefore, they can move the ball from a non-dangerous area to a medium-high dangerous place through valuable long ball passes.

Moreover, it is possible to apply the same visualization filtering by the team to analyze different players' on-ball performance levels within and across positions while focusing on a particular team, for example, understanding which central backs have the highest and least performance level concerning on-ball actions.

The following visualization compares the number of actions per 90 minutes with the average action value. Additionally, the highest rating player of each level of touches per 90 minutes is highlighted.

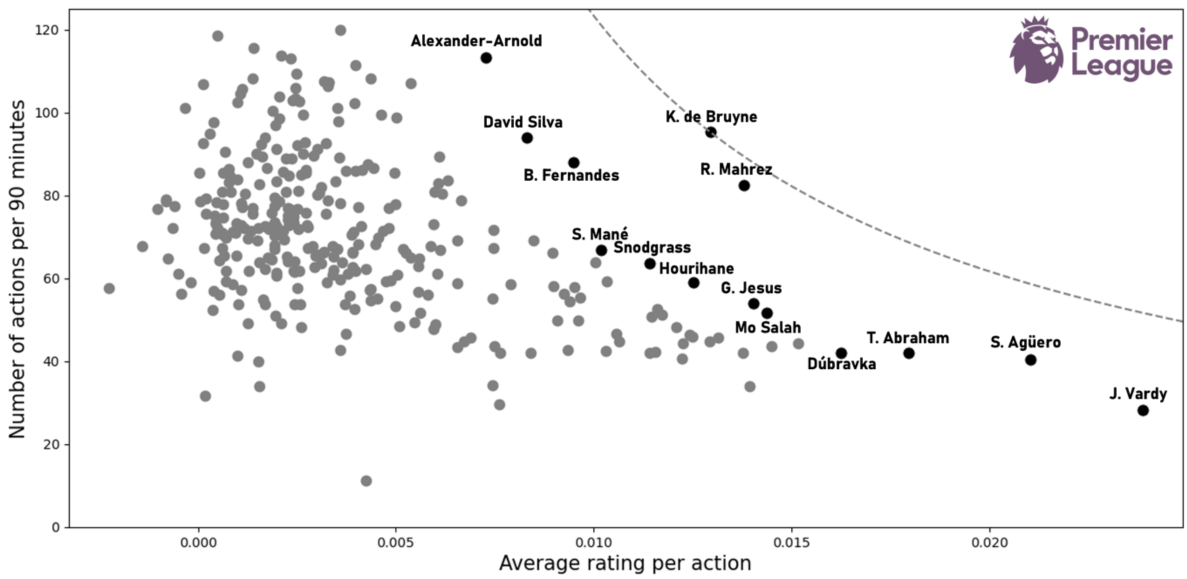


Figure 30 - Frequency vs Value

Usually, goalkeepers and forward players, such as Martin Dúbravka and Jamie Vardy, respectively, tend to perform fewer actions, whereas being able to impact the game. Conversely, game builders such as midfielders (e.g., David Silva and Bruno Fernandes) tend to participate in the game more often, therefore, regularly possess higher levels of touches per 90 minutes. However, in football, the difficulty resides in possessing high levels of on-ball touches and simultaneously producing valuable actions. Said that, from Figure 30, there is evidence that the Citizens, Kevin de Bruyne and Riyad Mahrez, are players capable of frequently performing valuable actions.

Additionally, the VAEP framework can be exploited as an indicator of on-ball performance. Therefore, it is feasible to produce player analysis to comprehend if players are under or overperforming in comparison to their standards or against players in league with similar roles while focusing on user-defined matches or even studying the entire season time window.

Concerning Manchester United Portuguese winter transfer, Bruno Fernandes (see Figure 31), his most frequent rating per game is approximately 0.5, representing a reference point to study Bruno's performance. In Gameweek 30 draw against Spurs, even though the Portuguese scored from the penalty spot, there is evidence that Bruno performed under his standard levels. On the bright side, Gameweek 33 home win versus Bournemouth, Bruno scored one goal and made two assists, directly impacting the Red Devils' final score.

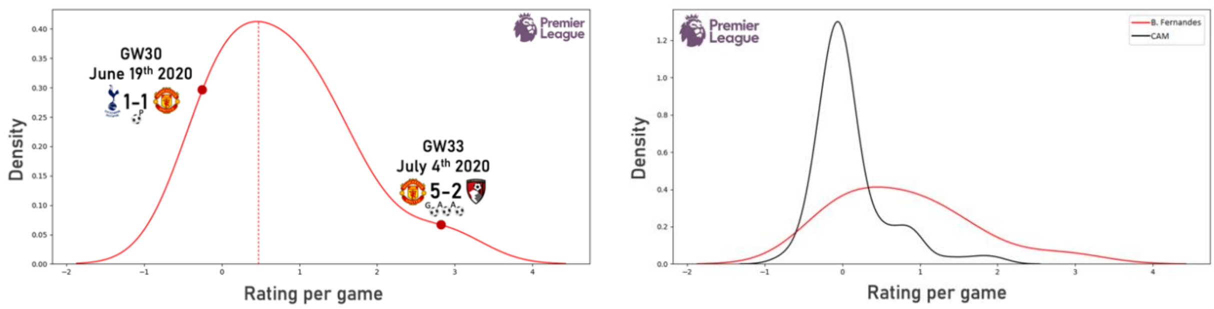


Figure 31 - Bruno Fernandes VAEP distribution

Furthermore, it is possible to change the granularity of players under the study to compare player performance levels against players in the league with similar roles or even analyze players within the same team. Figure 31 expresses the distribution of Bruno Fernandes in contrast to players in the league who share the same position, that is, Central Attacking Midfielder (CAM). As expected, the distribution of Bruno seems to be associated with higher ratings per game. In other words, in general, Bruno appears to have a higher performance level when compared to the overall Premier League attacking midfielders, according to the framework.

Throughout an entire season, is uncommon for players to perform in a single position, behavior that enlarges in more advanced positions on the pitch. Said that, through aggregating player action values by position, it is possible to perceive which positions allow players to be more impactful in the game.

Analyzing Citizen Bernardo Silva, there is evidence that Bernardo produces more valuable actions per ninety minutes when playing as Left Winger. However, it is the fourth option Pep Guardiola seems to assign to the Portuguese. In addition, typically, Bernardo tends to have a negative contribution to the game when playing as defensive midfielder, which complies with his physical characteristics (see Figure 32).

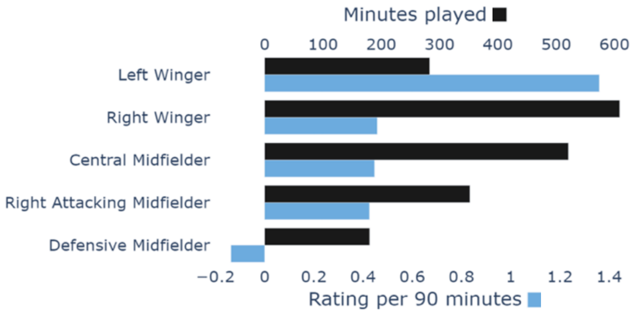


Figure 32 - Bernardo Silva performance by position

Another functionality of VAEP is to study the value of a user-defined action type in depth. In this sense, Figure 33 maps open play key passes by the Liverpool Right Back, Trent Alexander-Arnold, which are passes that result in a teammate shot, excluding assists.

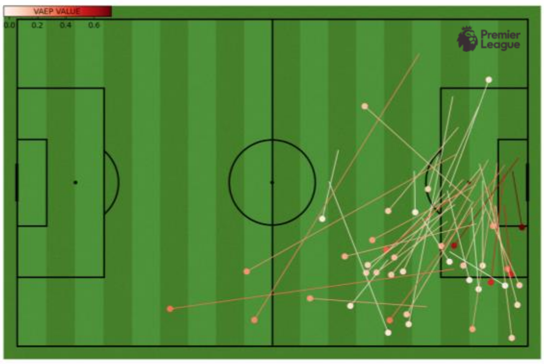


Figure 33 - Mapping Trent keypasses from open play

	PLAYER	TEAM	RATING /90	KEYPASS /90
1	K. de Bruyne	Man City	0.41	3.73
2	C. Hourihane	Aston Villa	0.31	1.93
3	R. Mahrez	Man City	0.31	2.37
4	Trent A-Arnold	Liverpool	0.29	2.10
5	P. Groß	Brighton	0.27	2.75

Table 6 - Top 5 keypass players

Due to Trent's position, there is a clear overflow through the right corridor, abounding actions at the entrance of the opponent box. Furthermore, there is a pattern where passes that end closer to the opponent goal seem more valuable to the framework, while backward passes tend to be associated with lower values. From Table 6, Trent Alexander-Arnold is the fourth most valuable player in this aspect of the game throughout the whole season, possessing a 0.29 rating per 90 minutes when exclusively focusing key passes, complemented by performing around two key passes per 90 minutes.

Nowadays, passing networks or maps are a standard methodology to study connections between players. These networks are constructed from observing the exchange of passes between pairs of teammates, where nodes (or vertices) represent the total number of completed passes each player performs, whereas links (or edges) account for the number of accurate passes each pair produces between them. Moreover, players are represented on the field considering the pass origin median location, while the color of players and connections represent the pass value according to the framework (see Figure 34).

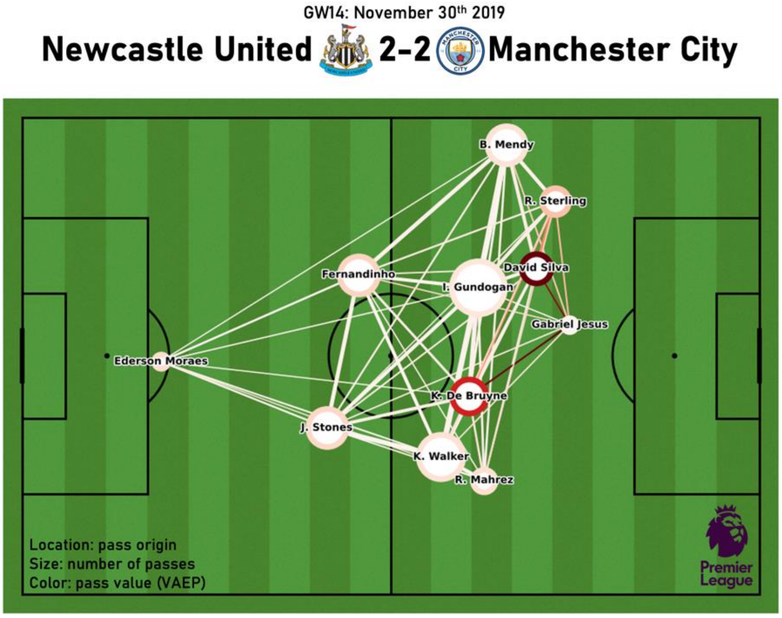


Figure 34 - Passing Network: Newcastle vs Manchester City

Focusing on Gameweek 14 two-all draw between Newcastle United and Manchester City, concerning team structure, there is an indication of possession in high block once the Citizens were collectively structured mainly in the opponent half. Individually, a notable distinction of positioning between both full-backs, Benjamin Mendy open on the line while Kyle Walker gives support from inside. In terms of pass frequency, Ilkay Gundogan was the main game builder performing 112 accurate passes, that is, 42 less than the whole squad of the Magpies. Notwithstanding, David Silva and Kevin de Bruyne were the players most involved in performing valuable passes. The Spaniard assisted Raheem Sterling and created a 1-on-1 goal opportunity against the goalkeeper for Gabriel Jesus. Besides scoring a goal of the season contender, the Belgian also created a clear goal opportunity for Raheem Sterling and possessed a strong connection with Gabriel Jesus. In support, the connections between offensive midfielders and Gabriel Jesus appear to be the most impactful links, followed by Raheem Sterling and both midfielders.

In the same sense, Figure 35 represents the value of actions performed by players when they were simultaneously on the pitch.

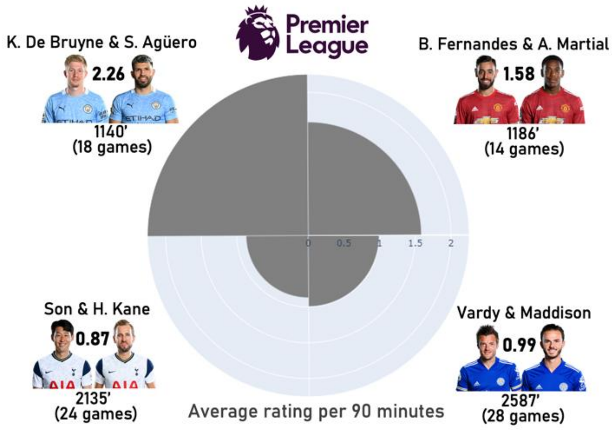


Figure 35 - Player connections (pairs)

By analyzing the performance of midfielders and forwards pairs over four teams that finished in position to access European competitions, Kevin de Bruyne and Sergio Aguero (2.26) outperformed the remaining pairs even though the Citizens shared the least amount of minutes played together (1140'). In sequence, Bruno Fernandes and Anthony Martial (1.58) also hold valuable figures nonetheless distant from Manchester rivals while sharing similar levels of minutes on the pitch. Moving to London, The Foxes shared the most minutes (2587'), while there is an indication that the Spurs were the least impactful pair out of the selection.

Once again, Figure 35 focuses on several teams across the league. However, depending on user needs, it is pleasant to filter by the team and study multiple combinations of players. For instance, study the performance of distinct midfield trios or defensive quartets within the same team.

In scouting, considering the hypothetical scenario where Manchester City desires to find potential replacements for David Silva, that is, identify players in the league with similar characteristics to the Spaniard.

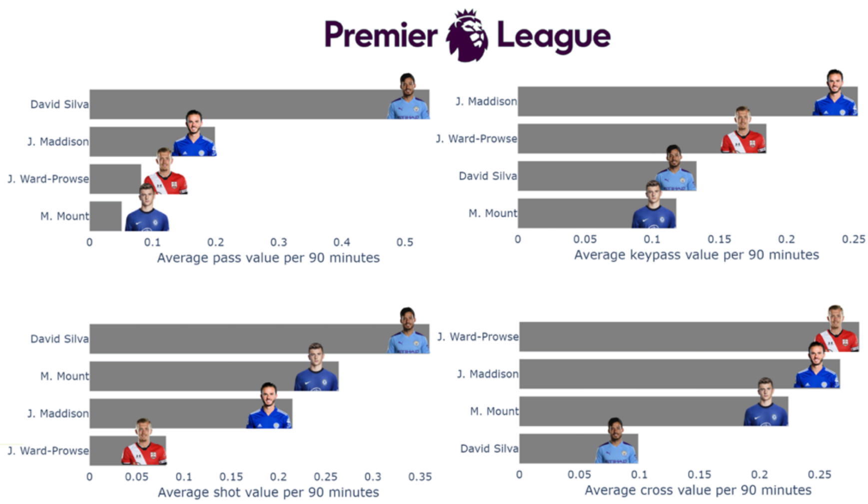


Figure 36 - Player similarity tool per action type

Regarding the most frequent action in football, David Silva (top 5 successful passes in opponent half per 90 minutes - 46) outperforms his competitors. However, in terms of key passes James Maddison (top 3 key passes - 76) and James Ward-Prowse (top 9 key passes - 59) were both able to possess more valuable figures than the four-time champion (top 28 key passes - 39).

Concerning shooting, Mason Mount (top 4 shots on target from midfielders - 29) is the player who becomes near David Silva levels, followed by James Maddison (top 4 shots from midfielders - 74).

There is evidence that crosses are not a valuable feature of the Spaniard. However, if the Citizens seek to acquire midfielders with such characteristics, possibly James Ward-Prowse (top 4 crosses completed from midfielders - 58) and James Maddison (top 3 crosses completed from midfielders - 60) are suitable targets.

Team Analysis

As previously identified, the framework can be exploited as an indicator of performance at the team level. In this sense, Figure 37 expresses the total value of Manchester City on-ball actions by game throughout the entire season. Once matches have approximately the same minutes, higher values could indicate higher performance levels.

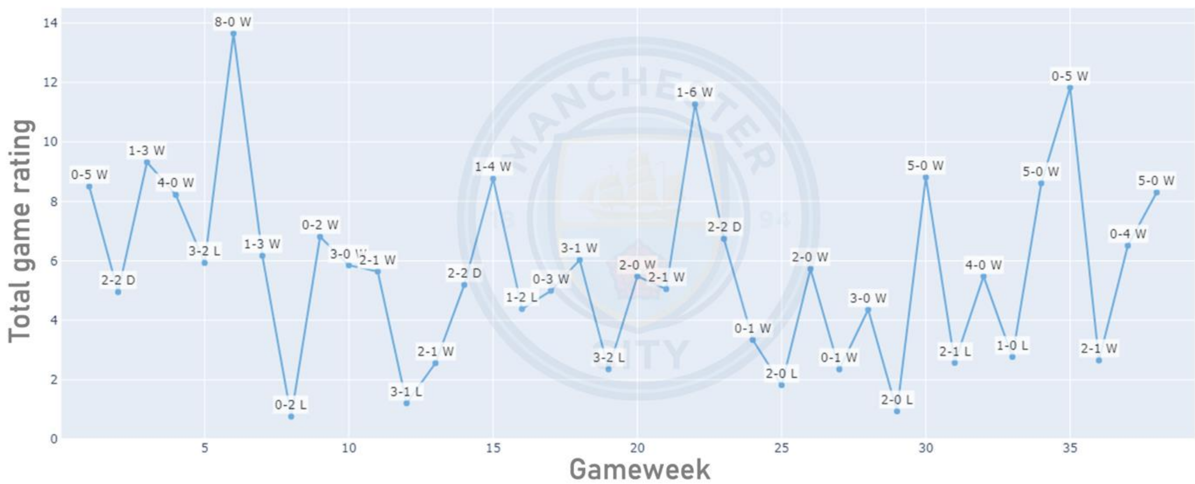


Figure 37 - Result vs. Performance: Manchester City

From analyzing Figure 37, it is feasible to identify high-performance moments, in concrete, 8-nil victory against Watford during game week 6 was the peak moment across the season. On the contrary, there is an indication that a two-nil loss against Wolves was the bottommost moment. Furthermore, the slope between two gameweeks gives a perception of how the team behaves compared to previous matches. In other words, comprehend if the performance level is increasing or decreasing over time. Also, Figure 37 compares the performance with the actual result to perceive if both are concordant, as an example, when the Citizens have a performance level similar to gameweek five lost against Norwich or gameweek 23 draw versus Crystal, typically the blues tend to win the game, indicating that the performance was better than the result. Moreover, it is possible to compare the team performance in specific matches against the team average level.

By interpreting the average game rating per team, it explains teams' actions on-ball value across the season. In this sense, Table 7 compares Premier League Table and VAEP Framework Ranking.

Premier League Table

	TEAM	POINTS
1	Liverpool	99
2	Man City	81
3	Man United	66
4	Chelsea	66
5	Leicester City	62
6	Tottenham	59
7	Wolves	59
8	Arsenal	56
9	Sheffield United	54
10	Burnley	54
11	Southampton	52
12	Everton	49
13	Newcastle	44
14	Crystal Palace	43
15	Brighton	41
16	West Ham	39
17	Aston Villa	35
18	Bournemouth	34
19	Watford	34
20	Norwich City	21

VAEP Framework Ranking

		TEAM	RATING
↑1	1	Man City	5.68
↓1	2	Liverpool	4.74
↑1	3	Chelsea	3.84
↑1	4	Leicester City	3.47
↓2	5	Man United	2.94
↑10	6	West Ham	2.93
↑10	7	Aston Villa	2.89
↓1	8	Wolves	2.84
↑1	9	Burnley	2.82
↑1	10	Southampton	2.72
↓3	11	Arsenal	2.67
↓6	12	Tottenham	2.61
↓1	13	Everton	2.59
↓5	14	Sheffield United	2.58
•	15	Brighton	2.39
↓3	16	Newcastle	2.38
↑1	17	Bournemouth	2.19
↓4	18	Crystal Palace	1.91
•	19	Watford	1.58
•	20	Norwich City	1.40

Table 7 - League Standings vs Framework Ranking

Notwithstanding, teams with higher possession values naturally have more opportunities to create valuable actions that do not necessarily translate into superiority. For example, during gameweek 8 Wolves won two-nil versus the Citizens even though the blues had over 75% possession.

From Table 7, it is meaningful to compare points obtained against teams' overall performance indicator. Man City rose to the top of the table for the mentioned reasons, overtaking the Reds. Additionally, there is evidence that West Ham and Aston Villa's performances appear to be more valuable than the results obtained once both teams climbed ten positions. In the opposite direction, Tottenham and Sheffield United lost six and five positions, respectively, indicating that results were better than on-ball performance levels. In the relegation zone, the framework indicates that Norwich and Watford deserved the Championship relegation while Crystal Palace did not present a performance index worthy of Premier League maintenance.

In addition, analyzing the standard deviation of team ratings per game provides a sense of consistency indicator across the season. Hence, it is possible to identify teams that tend to conserve a stable performance level and recognize teams characterized by volatile performances.

Match Analysis

The framework can be exploited to understand teams' momentum throughout user-defined matches. To do so, the game was divided into twenty equal bins of five minutes each, which was measured the sum of every on-ball action per team. Furthermore, once the main goal is to represent teams' dominance, the visualization focuses on the offensive component of the framework as performing valuable defensive actions might be associated with valuable offensive actions by the opponent, for example, shot taken by the striker against goalkeeper save.

Figure 38 illustrates the clash of Portuguese managers between José Mourinho and Nuno Espírito Santo in gameweek 28.

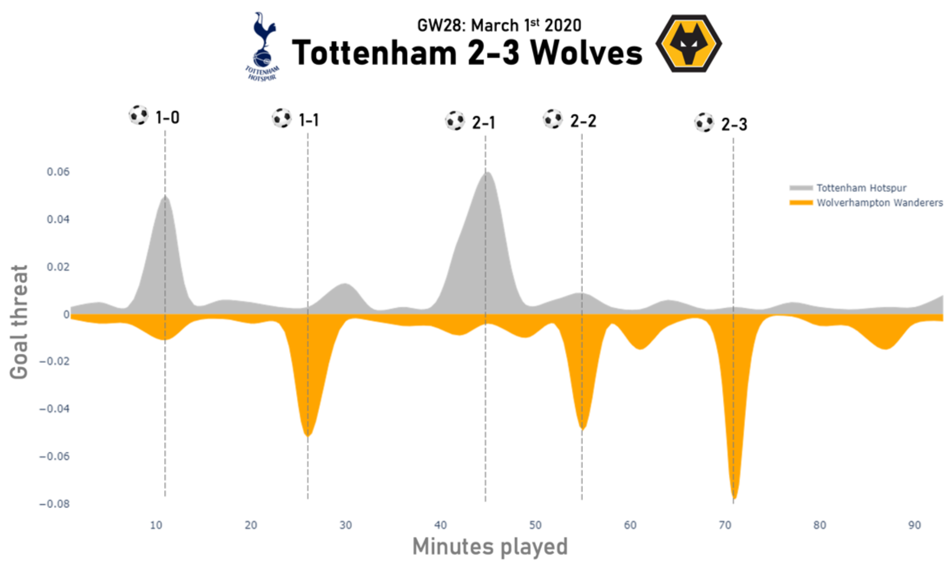


Figure 38 - Match momentum: Tottenham vs Wolverhampton

The first half was characterized by a team scoring and the opponent's reaction, finishing with Spurs' dominance near the first-half whistle. In the second half, the Wolves had the majority dominance of the game in creating goal opportunities hence turning the result around. In addition, it is meaningful to exclude goal contributions to possess a deeper understating of teams' behavior and identify cases of creating momentum towards the goal as opposed to scoring during a period where the opponent was theoretically dominating the game.

4.2.8. Discussion

This section will interpret the main achievements and highlight the major strengths, shortcomings, and conceivable improvements for future work.

The major achievement of VAEP over most existing works resides in valuing all action types (e.g., passes, shots, clearances, and tackles), allowing to convert everything that happens in the beautiful game to a single currency. Furthermore, the framework quantifies offensive and defensive on-ball contributions. Also, as the code was implemented from scratch, it is completely maneuverable and flexible in assembling alterations or even reinforcing the current approach.

In addition, contrary to previous studies, the methodology adopted is non-exclusively goal-dependent by adding Expected Goals (xG) metric during the characterization of the target variables, once goals are a very rare occasion of the game and directly dependent on the attacker's finishing ability. Notwithstanding, there is awareness concerning xG metric denoting a shot-dependent method. That is, xG disregards obvious goal opportunities where the attacker concluding the sequence of play is unable to shoot. Furthermore, game state representation contains information about the short past while considering an action's possible effects on subsequent actions.

Finally, the model bases its valuation on the game context in which actions occur. That is, the goal difference before an action occurs once teams often adapt their playing style according to the current scoreline (Decroos et al., 2019). Moreover, the model is conscious of the moment in which actions happen, i.e., actions performed during open play instead of set pieces. This method makes it scalable for teams to use as focuses on event data whereas tracking systems can be expensive.

Football analytics has long focused on the outcomes of discrete on-ball events. However, much of the sport's complexity resides in off-ball circumstances. In the words of Johan Cruyff: "it is statistically proven that players have the ball 3 minutes on average. So, the most important thing is: what do you do during those 87 minutes when you do not have the ball? That is what determines whether you are a good player or not." (Fernández, 2018). Hence, the main deficiency of the current approach resides in disregarding off-ball features.

<u>Disregard off-ball movements (offensively and defensively)</u>

As an illustrative offensive example, Figure 39 focuses N'Golo Kanté's off-ball movement during the 93rd minute of a Champions League round of 16. Even though the Frenchman has not performed any on-ball action across his 90-meter dummy run, Kanté created a three-versus-one goal opportunity. Emerson was able to materialize and eliminate Atlético de Madrid.

Figure 39 - N'Golo Kanté valuable offensive off-ball movement

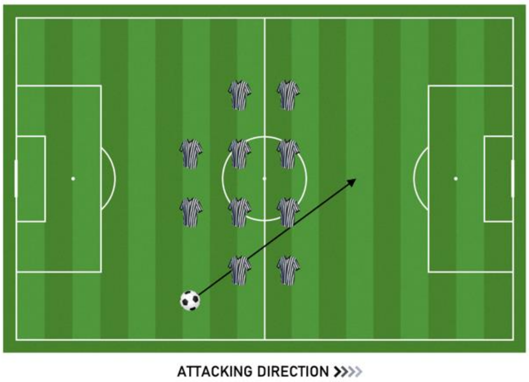
Virgil van Dijk faced a two-versus-one situation against Heung-min Son and Moussa Sissoko from a defensive perspective. Once more, the center-back has not taken any touch on the ball. Nonetheless, through Virgil's positioning and defensive containment, the Dutchman forced his opponent to act deficiently and waste a precise goal occasion, Figure 40.

Figure 40 - Virgil Van Dijk valuable defensive off-ball movement

Since players did not execute any action on the ball, the framework omits these movements in both cases. Hence, the current approach disregards off-ball contributions.

<u>Ignore player's off-ball positioning (opponents and teammates)</u>

While studying the opponent's off-ball positioning, Figures 41 and 42 emphasize two identical passes which share the same feature values (i.e., location, successfulness, distance covered, angle, spin, et cetera). However, in scenario A the opponent holds eleven men behind the ball line in contrast to scenario B, where the opponent possesses all outfield players ahead of the pass end location. In this case, the opponent's defensive structure influences the importance of the action. Nevertheless, the framework does not capture this information.



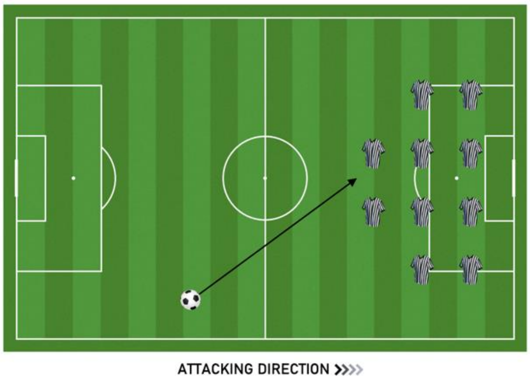
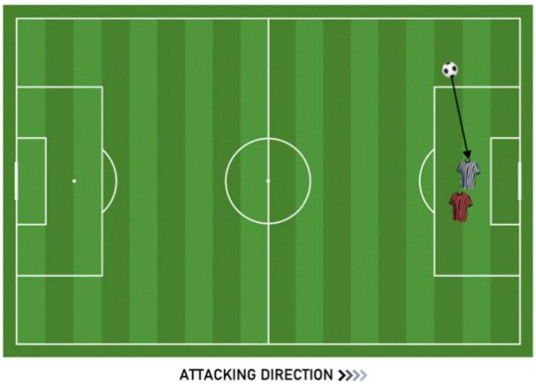


Figure 41 - Scenario A: Opponent's defensive structure

Figure 42 - Scenario B: Opponent's defensive structure

Following the same reasoning, the framework also disregards teammates off-ball positioning and opponents. In this sense, Figures 43 and 44 represent two equal crosses, in which scenario C illustrates an opponent behind the last defender. In contrast, in scenario D, the defensive line has control of the box. Under these circumstances, assuming that the defender performs the exact clearance inside the box in both scenarios, the clearance performed under the first scenario embodies a more valuable action once the defender denied an opponent's scoring opportunity, opposed to scenario D where taking into consideration pitch control, ideally, the defender could adopt a more conservative action while retaining possession.



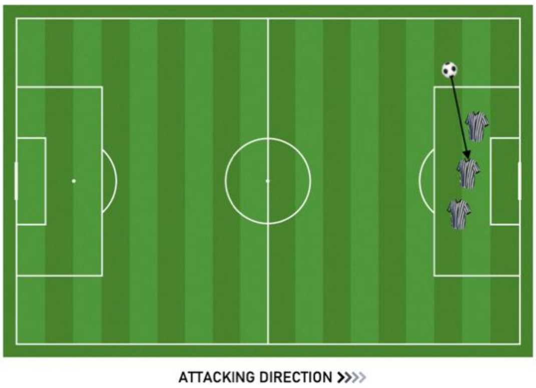


Figure 43 - Scenario C: Opponent's off-ball positioning

Figure 44 - Scenario D: Teammate's off-ball positioning

<u>Ignore on-ball pressure</u>

In football, most actions are dependent on the context in which they occur. Said that, Figures 45 and 46 ensembles alike shots which vary the on-ball pressure imposed by the opponent. In this regard, scenario E exemplifies four opponents close to the ball in contrast with scenario F, where a single opponent is applying pressure. In this case, scenario F is more likely to result in a successful shot considering the opponent's positioning. However, the current framework does not take into consideration on-ball pressure.

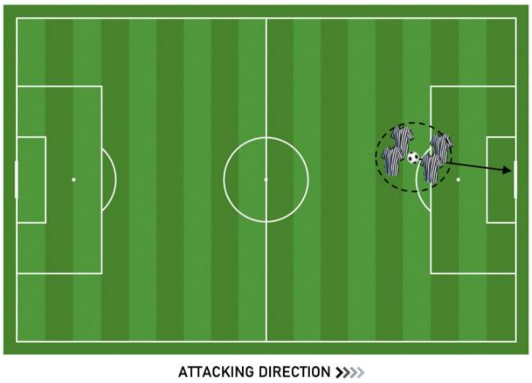


Figure 45 - Scenario E: Opponent's on-ball pressure

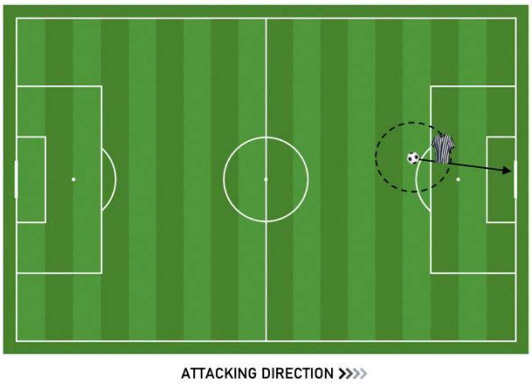


Figure 46 - Scenario F: Opponent's on-ball pressure

Defensive players are not penalized for allowing opposition attacking actions to occur

In the same sense, defensive players are not penalized for allowing opposition attacking actions to occur. As a representative example, Figure 47 expresses Wolverhampton's shooting creation against Manchester United. Wolves managed to find space at the entrance of the box for Rúben Neves to shoot and score a stunning goal due to a studied corner. In such circumstances, the Red Devils must be aware of the Portuguese midfielder's long-range finishing capabilities, not allowing Neves to shoot from that area comfortably.

Figure 47 - Video representation of Rúben Neves mid-range free shot

Action body part unknown

According to previous studies, the body part in which players perform actions can influence their outcome. Therefore, there is proof of the relevance of such variables to determine the value of actions. Despite such evidence from the data provider, the body part feature contains more than 90% missing values. Consequently, this feature was excluded from the analysis due to the lack of information.

Conversely, the framework also has a limitation for being able to handle progressive actions that are focused on keeping possession instead of advancing closer to the goal.

In sum, there is evidence that, when available, adding tracking data that captures players' spatial location by timeframe in future work will provide valuable context to actions, therefore, approximating the quantification to the game and capturing behaviors that the current approach is lacking through positional play. Finally, Table 8 summarizes the identified strengths and shortcomings of the developed framework.

STRENGTHS	SHORTCOMINGS
Credit all on-ball action types (e.g. passes, shots, tackles)	Disregard off-ball movements (offensively and defensively)
Non exclusively goal dependent approach	Ignore player's off-ball positioning
Captures short-past information	Ignore on-ball pressure
Captures action's effect on subsequente actions	Defensive players are not penalized for allowing opposition attacking actions to occur
Features contextual play information	Action body part unknown

Table 8 - Summary of VAEP Framework strengths and shortcomings

“If what you did yesterday seems big, you haven’t done anything today.”
- Lou Holtz

5. LIMITATIONS AND FUTURE WORK

The data-driven tools developed under this report were fully accomplished on top of event data powered by Opta. In this regard, the major limitation lies in lacking teams' positional structure, capturing the environment around the ball, and quantifying tactical and physical efforts from tracking data.

In addition, Pass Sonar and VAEP projects focus on player-level and action-level technical analysis, respectively, to facilitate performance-related data storage, management, analysis and communication across the club. Nonetheless, football is such a diverse sport with numerous tactics, unique scenarios, and systems of playing. Hence, the next step embodies team-level tactical analysis, which aims to characterize team playing styles by considering multiple events and factors that determine a team's playing style. The strength of this approach is represented by a membership value between 0% and 100% assigned to every team possession, based on the individual definitions and calculations. To capture the different ways a team plays, an algorithm widely utilized in Natural Language Processing (NLP), especially within topic modelling, is employed, Latent Dirichlet Allocation (LDA).

Concretely, topic modelling refers to discovering abstract topics in a collection of documents. In addition, documents are probability distributions over latent topics, whereas topics are probability distributions over words. In the analogous sense, topics in language analysis resemble play styles in football data analysis. Hence, the goal involves identifying the mixture of playing styles in a collection of matches per team. This implementation is adapted from Team Personas (Meza et al., 2018). Notwithstanding, the framework differentiates the definitions of playing styles studied and their implications, as shown in Table 9.

Playing style	Description
Maintenance	Possessions in which teams look to maintain and secure possession of the ball within the defensive half of the pitch.
Build-up	Possessions in which teams look for opportunities to attack within the first 25% of the offensive half of the pitch.
Sustained Threat	Possessions in which teams look to create a goal opportunity within the offensive third of the pitch.
High Pressure	Captures teams' regaining possession in the last 40% of the pitch.
Crossing	It occurs if the ball is delivered from a wide area of the pitch to find a teammate inside the opponent's box.
Fastbreaks	It occurs once a team regains possession and moves the ball into an attacking area via passes, dribbles, or a combination of both.
Direct Play	Captures instances of play in which teams move the ball quickly towards the opposition's goal through long passes rather than sustained possession.
Dogged Defending	Features active defending teams in the defensive half of the pitch.

Table 9 - Definition of playing styles

Thus, in language analysis, LDA receives a word frequency matrix as input, containing the number of times a word appeared in a document. Similarly, in football analysis, words are replaced by actions using event data. In this sense, the algorithm contains information about the number of times a set of features occurred during a 90-minute match. As a result, the algorithm provides a probability inherent to each component that characterizes play styles.

As a representative illustration, Pep’s dominance through dynamic possession style can be observed throughout the two-nil defeat at Etihad Stadium, counting towards gameweek 8 (see Figure 48).

Figure 48 - Manchester City’s playing style against Wolves

In addition, averaging all games across a season-long time window characterizes teams' playing styles. As can be seen from Table 10, Man City is denoted by playing possession in the opponent's half (i.e., build-up and sustained threat styles). Therefore, this game control behavior allows the runners-up to remain inactive defensively, featuring the lower value across dogged defending style in the league.

In contrast, there is an indication that the Wolves spend most of the games in their defensive midfield under Nuno Espírito Santo, by playing in low block defensively and starting to build from the back, which reflects high values of dogged defending and maintenance, respectively. Moreover, Nuno's side represents the squad with the highest fastbreaks' membership value, which suits the characteristics of Adama Traoré and Pedro Neto in offensive transition moments.

Lastly, Burnley is the second busiest team defensively, only surpassed by Newcastle, while focusing their attacking outcome on direct play through targeting Chris Wood's physical characteristics.

Team	Maintenance	Build-up	Sustained Threat	High Pressure	Crossing	Fastbreaks	Direct Play	Dogged Defending
Manchester City	0.14	0.31	0.20	0.05	0.21	0.03	0.03	0.03
Liverpool	0.17	0.23	0.14	0.10	0.15	0.06	0.09	0.06
Chelsea	0.25	0.17	0.13	0.09	0.18	0.06	0.04	0.07
Arsenal	0.33	0.07	0.11	0.08	0.13	0.08	0.07	0.11
Manchester United	0.18	0.21	0.13	0.07	0.16	0.08	0.05	0.13
Leicester City	0.22	0.15	0.11	0.08	0.14	0.10	0.06	0.14
Brighton	0.28	0.08	0.12	0.08	0.11	0.07	0.10	0.16
Everton	0.20	0.07	0.09	0.10	0.08	0.10	0.16	0.18
Tottenham	0.22	0.19	0.07	0.08	0.09	0.10	0.05	0.20
Norwich City	0.33	0.11	0.06	0.08	0.09	0.07	0.06	0.21
Southampton	0.10	0.13	0.08	0.10	0.07	0.07	0.23	0.22
Wolves	0.20	0.10	0.08	0.08	0.09	0.14	0.09	0.23
West Ham	0.09	0.16	0.08	0.09	0.06	0.12	0.16	0.23
Sheffield United	0.06	0.09	0.19	0.05	0.05	0.10	0.23	0.23
Bournemouth	0.18	0.07	0.13	0.07	0.07	0.08	0.15	0.24
Aston Villa	0.14	0.07	0.11	0.08	0.07	0.12	0.16	0.25
Watford	0.08	0.10	0.13	0.08	0.06	0.09	0.21	0.26
Crystal Palace	0.18	0.08	0.10	0.09	0.07	0.08	0.13	0.27
Burnley	0.08	0.11	0.05	0.06	0.04	0.08	0.30	0.27
Newcastle United	0.17	0.07	0.06	0.07	0.06	0.12	0.10	0.34

Table 10 - Characterization of teams' playing style across 2019/20 Premier League season

In this manner, the following practical questions can be exploited:

1) What’s the playing style that characterizes a team?
2) Is the team playing style consistent over different competitions?
3) Does the playing style change over different formations?
4) What’s the impact of changing coach on the style of play?
5) Is the teams’ playing style consistent across several seasons?
6) What’s the most effective playing style to play against a certain team’s playing style?
7) What’s the most suitable playing style for players?
8) What playing style allows players to perform more valuable actions?
9) Against what playing style do certain players perform better?

Notwithstanding, similar to many other tactical elements in football, this data-driven methodology contains a certain lack of precision, in the sense that a clear ground truth reference for the accuracy of the measurement cannot exist.

“Not everything that can be counted counts and not everything that counts can be counted.”

- Albert Einstein

6. CONCLUSION

In an open nature sport characterized by unique scenarios, various game contexts, and arbitrary behaviors by players, the analysis of player contributions through technical on-ball actions becomes a critical element for understanding performance. According to Pep Guardiola, the satisfaction taken from results lasts five minutes, but what remains sustainable and gives confidence for the future is the performance.

Beyond the bigger picture that overall performance statistics of multiple matches can provide, the understanding of on-ball actions demands the need for a more specialized data-centric solution. In this regard, the developed model that captures a comprehensive set of contextual features shows how it is possible to extract meaningful information by converting everything that happens in the beautiful game to a single currency through leveraging event data. Hence, technical teams can provide specific guidance to improve individual-player and team-collective performances by understanding players' contributions' frequencies, quality, and effectiveness.

Furthermore, representing the frequency and distance of the most recurrent technical action in the game by direction (i.e., passes) allows identifying possession patterns at the player level. With this tool, it is achievable to assess players' passing trends, identify game builders, infer high-pressure triggers, visualize the impact of substitute players, and recognize unsuccessful passing tendencies across different moments in the game.

Finally, this report showcases technical analysis to support decision-making within football associations. Beyond that, tactical analysis exploiting Natural Language Processing (NLP) to characterize teams' playing styles can reinforce the presented work by treating coach language as a bag-of-words model, where each playing style is a word and each game represents a collection of words, analogously to sentences. Thus, the NLP methodology addressed shows how to bridge the gap between industry-specific language and data.

7. BIBLIOGRAPHY

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: A Parallel Overview.

Beckmann, M., Ebecken, N. F. F., & Pires de Lima, B. S. L. (2015). A KNN Undersampling Approach for Data Balancing. *Journal of Intelligent Learning Systems and Applications*, *07*(04), 104–116. https://doi.org/10.4236/jilsa.2015.74010

Bishop, C. M. (1995) *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford.

Breiman, L. (2001). Random Forests (Vol. 45).

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446–3453. https://doi.org/10.1016/j.eswa.2011.09.033

Buskirk, T. D. (2018). Surveying the Forests and Sampling the Trees: An overview of Classification and Regression Trees and Random Forests with applications in Survey Research. *Survey Practice*, *11*(1), 1–13. https://doi.org/10.29115/sp-2018-0003

Cervone, D., D'Amour, A., Bornn, L., & Goldsberry, K. (2016). A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes. *Journal of the American Statistical Association*, *111*(514), 585–599. https://doi.org/10.1080/01621459.2016.1141685

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-setp data mining guide. *SPSS inc, 9,* 13.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*, 785–794. https://doi.org/10.1145/2939672.2939785

Davenport, T. 2014. Analytics in sports: The new science of winning. *International Institute for Analytics*, 1-28.

Decroos, T., van Haaren, J., Bransen, L., & Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1851–1861. https://doi.org/10.1145/3292500.3330758

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), 64–73. https://doi.org/10.1145/2500499

Donoho, D. (2017). 50 Years of Data Science. In *Journal of Computational and Graphical Statistics* (Vol. 26, Issue 4, pp. 745–766). American Statistical Association. https://doi.org/10.1080/10618600.2017.1384734

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, *35*(5–6), 352–359. https://doi.org/10.1016/S1532-0464(03)00034-0

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine,* 17(3), 37-54.

Fernández, J., Barcelona, F. C., Bornn, L., Los, D. C., & Dodgers, A. (2019). Decomposing the Immeasurable Sport: A deep learning expected possession value framework for soccer. https://lukebornn.com/sloan_epv_curve.mp4

Fernández, J., Barcelona, F. C., Fernandez, J., & Bornn, L. (2018). Wide Open Spaces: A statistical technique for measuring space creation in professional soccer. http://www.lukebornn.com/sloan/space_occupation_1.mp4

Ghawi, R., & Pfeffer, J. (2019). Efficient Hyperparameter Tuning with Grid Search for Text Categorization using kNN Approach with BM25 Similarity. *Open Computer Science*, *9*(1), 160–180. https://doi.org/10.1515/comp-2019-0011

Hughes, M., Caudrelier, T., James, N., Redwood-Brown, A., Donnelly, I., Kirkbride, A., & Duschesne, C. (2012). Moneyball and soccer - An analysis of the key performance indicators of elite male soccer players by position. *Journal of Human Sport and Exercise*, *7*(SPECIALISSUE.2), 402–412. https://doi.org/10.4100/jhse.2012.72.06

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study 1. In *Intelligent Data Analysis* (Vol. 6). IOS Press.

Kempton, T., Kennedy, N., & Coutts, A. J. (2016). The expected value of possession in professional rugby league match-play. *Journal of Sports Sciences*, *34*(7), 645–650. https://doi.org/10.1080/02640414.2015.1066511

Khalid, S. (2014). A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning (Vol. 372). www.conference.thesai.org

Kleinbaum, D. G., & Klein, M. (2010). Logistic Regression A Self-Learning Text Third Edition. http://www.springer.com/series/2848

Kotu, V., & Deshpande, B. (2019). Data Science Process. In *Data Science* (pp. 19–37). Elsevier. https://doi.org/10.1016/b978-0-12-814761-0.00002-2

Kriesel, D. (2007). *A Brief Introduction to Neural Networks.*

Kubiak, T., & Benbow, D. W. (2005). *The certified six sigma black belt handbook.* Milwaukee, WI: ASQQuality Press.

Leinweber, D. J. (2007). Stupid Data Miner Tricks. *The Journal of Investing.* https://doi.org/10.3905/joi.2007.681820

Liu, G., & Schulte, O. (2018). Deep Reinforcement Learning in Ice Hockey for Context-Aware Player Evaluation. https://doi.org/10.24963/ijcai.2018/478

Mackay, N. 2017. Improving my ‘xG added’ model. https://mackayanalytics.nl/2017/07/28/improving-my-xg-added-model

Malekian, A., & Chitsaz, N. (2021). Concepts, procedures, and applications of artificial neural network models in streamflow forecasting. *Advances in Streamflow Forecasting*, 115–147. https://doi.org/10.1016/B978-0-12-820673-7.00003-2

Marbán, Ó., Mariscal, G., & Segovia, J. (2009). A data mining & knowledge discovery process model. Data mining and knowledge discovery in real life applications. IntechOpen.

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2019). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

McHale, I. G., Scarf, P. A., & Folker, D. E. (2012). On the development of a soccer player performance rating system for the english Premier League. *Interfaces*, *42*(4), 339–351. https://doi.org/10.1287/inte.1110.0589

McHale, I., & Scarf, P. (2007). Modelling soccer matches using bivariate discrete distributions with general dependence structure. In *Statistica Neerlandica* (Vol. 61, Issue 4).

Miller, G. W. (2020). Data science and the exposome. In *The Exposome* (pp. 181–209). Elsevier. https://doi.org/10.1016/b978-0-12-814079-6.00008-0

Pappalardo, L., Cintia, P., Ferragina, P., Massucco, E., Pedreschi, D., & Giannotti, F. (2019). PlayeRank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technology*, *10*(5). https://doi.org/10.1145/3343172

Park, Y. S., & Lek, S. (2016). Artificial Neural Networks: Multilayer Perceptron for Ecological Modeling. In *Developments in Environmental Modelling* (Vol. 28, pp. 123–140). Elsevier B.V. https://doi.org/10.1016/B978-0-444-63623-2.00007-4

Pollard, R. (2004). Estimating the probability of a shot resulting in a goal: The effects of distance, angle and space. https://www.researchgate.net/publication/240641737

Romer, D. (2006). Do firms maximize? Evidence from professional football. *Journal of Political Economy*, *114*(2), 340–365. https://doi.org/10.1086/501171

Routley, K. (2015). A Markov Game Model for Valuing Player Actions in Ice Hockey.

Roy, M. van, Robberechts, P., Decroos, T., & Davis, J. (2020). Valuing On-the-Ball Actions in Soccer: A Critical Comparison of xT and VAEP. www.aaai.org

Rudd, S. 2011. A Framework for Tactical Analysis and Individual Offensive Production Assessment in Soccer Using Markov Chains. In *New England Symposium on Statistics in Sports.*

Rumelhart, D. E., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations* (pp. 318- 362). Cambridge, MA: MIT Press.

Schapire, R. E. (2003). The Boosting Approach to Machine Learning: An Overview.

Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA) Data Mining in Healthcare for Heart Diseases View project A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA) View project A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). In *International Journal of Innovation and Scientific Research* (Vol. 12, Issue 1). http://www.ijisr.issr-journals.org/

Singh, K. 2019. Introducing expected threat. https://karun.in/blog/expected-threat.html

Spearman, W. (2018). Beyond Expected Goals. https://www.researchgate.net/publication/327139841

Sutton, C. D. (2005). Classification and Regression Trees, Bagging, and Boosting. In *Handbook of Statistics* (Vol. 24, pp. 303–329). Elsevier. https://doi.org/10.1016/S0169-7161(04)24011-1

Tambe, P., & Hitt, L. M. (2012). The productivity of information technology investments: New evidence from IT labor data. *Information Systems Research*, *23*(3 PART 1), 599–617. https://doi.org/10.1287/isre.1110.0398

Thakkar, P., & Shah, M. (2021). An Assessment of Football Through the Lens of Data Science. In *Annals of Data Science* (Vol. 8, Issue 4, pp. 823–836). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s40745-021-00323-2

Tu, J. v. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, *49*(11), 1225–1231. https://doi.org/10.1016/S0895-4356(96)00002-9

Van der Aalst, W. (2016). Data Science in Action. In *Process Mining* (pp. 3–23). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-49851-4_1

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining.

Yam, D. 2019. Attacking contributions: Markov models for football. https://statsbomb.com/2019/02/attacking-contributions-markov-models-for-football/

Zambom-Ferraresi, F., Rios, V., & Lera-López, F. (2018). Determinants of sport performance in European football: What can we learn from the data? *Decision Support Systems*, *114*, 18–28. https://doi.org/10.1016/j.dss.2018.08.006

8. APPENDIX

Variable Name	Description	Variable Type	Example Data
Player Name	Player who performed the pass	Nominal variable	Rúben Neves; Kevin de Bruyne
Shirt Number	Player’s shirt number	Ordinal variable	1; 10
Position	Player’s starting position	Nominal variable	Goalkeeper; Centre Forward
Club	Player’s team name	Nominal variable	Aston Villa; Manchester United
Season	Game season	Nominal variable	2019/2020
Competition	Game competition	Nominal variable	Premier League
Game Week	Game week	Ordinal variable	1; 34
Formation	Team formation style	Nominal variable	4_4_2; 4_3_3
Home	Home team identifier	Binary variable	0; 1
Pass Angle	Pass angle in degrees	Continuous variable	6; 350
Pass Length	Pass length in meters	Continuous variable	10.8; 32.2
x	x coordinate of the player's starting position	Continuous variable	6.1; 43.6
y	y coordinate of the player's starting position	Continuous variable	12.8; 58.1
Success	Successful pass identifier	Binary variable	0; 1
Period	Game period	Ordinal variable	1; 2
Game Clock	Moment when the pass occurred in seconds	Continuous variable	3; 5684
Sub In	Second the substitute player entered the game	Continuous variable	0; 2700

Table 11 - Pass Sonar metadata

Variable Name	Description	Variable type
Pass	Action type	Binary variable
Crosses	Action type	Binary variable
Shot	Action type	Binary variable
Shot_On_Target	Action type	Binary variable
Shot_OffTarget	Action type	Binary variable
Shot_Blocked	Action type	Binary variable
Recovery	Action type	Binary variable
ChallengeLost	Action type	Binary variable
TakeOn	Action type	Binary variable
TurnOver	Action type	Binary variable
Tackle	Action type	Binary variable
Clearance	Action type	Binary variable
Interception	Action type	Binary variable
Aerial	Action type	Binary variable
BlockedCross	Action type	Binary variable
BadTouch	Action type	Binary variable
BlockPass	Action type	Binary variable
Disposs	Action type	Binary variable
Freekick	Set piece action type	Binary variable
Corner	Set piece action type	Binary variable
ThrowIn	Set piece action type	Binary variable
GKSave	Goalkeeper action type	Binary variable
GK_CrossClaim	Goalkeeper action type	Binary variable
GKPunch	Goalkeeper action type	Binary variable
GKThrows	Goalkeeper action type	Binary variable
GKPickup	Goalkeeper action type	Binary variable
Foul	Action result	Nominal variable

Offside	Action result	Binary variable
KeyPass	Action result	Binary variable
Goal	Action result	Binary variable
OwnGoal	Action result	Binary variable
Assist	Action result	Binary variable
Penalti	Action result	Binary variable
Post	Action result	Binary variable
Booking	Action result	Ordinal variable
Success	Action result	Binary variable
possession_loss	Action result	Binary variable
currentScoreDiff	Goal difference before the action	Discrete variable
start_x	Start location of the action in the x direction	Continuous variable
start_y	Start location of the action in the y direction	Continuous variable
end_x	End location of the action in the x direction	Continuous variable
end_y	Start location of the action in the y direction	Continuous variable
x_distance	Distance covered during the action in x direction	Continuous variable
y_distance	Distance covered during the action in y direction	Continuous variable
distance_covered	Distance covered during the action in meters	Continuous variable
start_distance_to_goal	Start euclidean distance to goal	Continuous variable
end_distance_to_goal	End euclidean distance to goal	Continuous variable
start_angle_to_goal	Start angle to offensive goal	Continuous variable
end_angle_to_goal	End angle to offensive goal	Continuous variable

offtarget_distance_to_goal	Off-target shot euclidean distance to center of the goal	Continuous variable
PassAngle	Pass angle (0 to 360 degrees)	Continuous variable
GoalLabel	(=1) if the team possessing the ball scored a goal or take a shot with xG > threshold, in the subsequent k actions (k=10), (= 0) otherwise	Binary target variable
ConcedeLabel	(=1) if the team possessing the ball conceded a goal or a shot with xG > threshold, in the subsequent k actions (k=10), (= 0) otherwise	Binary target variable

Table 12 - VAEP metadata

Model	Parameter	Value range	Optimal value
XGBoost	n_estimators	[50, 100, 200]	50
	max_depth	[3, 5, 10]	5
	learning_rate	[0.01, 0.1, 0.3, 0.5]	0.3
Neural Network	hidden_layer_sizes	[(100, 100, 100), (100, 100), (100), (50), (50, 50), (50, 50, 50)]	(100)
	activation	[identity, logistic, tanh, relu]	logistic
	learning_rate	[constant, adaptive]	constant
	solver	[lbfgs, sgd, adam]	lbfgs
	alpha	[0.0001, 0.001, 0.01]	0.0001
	learning_rate_init	[0.001, 0.01, 0.1]	0.001
	batch_size	[64, 128, 200, 512, auto]	auto
Random Forest	n_estimators	[100, 200, 300]	N/A
	max_depth	[None, 10, 50, 100]	N/A
	criterion	[gini, entropy]	N/A
	warm_start	[True, False]	N/A
	class_weight	[None, {0: 0.1, 1: 0.9}, {0: 0.4, 1: 0.6}, balanced]	N/A
Logistic Regression	solver	[newton-cg, sag, saga, lbfgs]	newton-cg
	warm_start	[True, False]	True
	class_weight	[None, {0: 0.1, 1: 0.9}, {0: 0.4, 1: 0.6}, balanced]	None

Table 13 - Parameters grid space

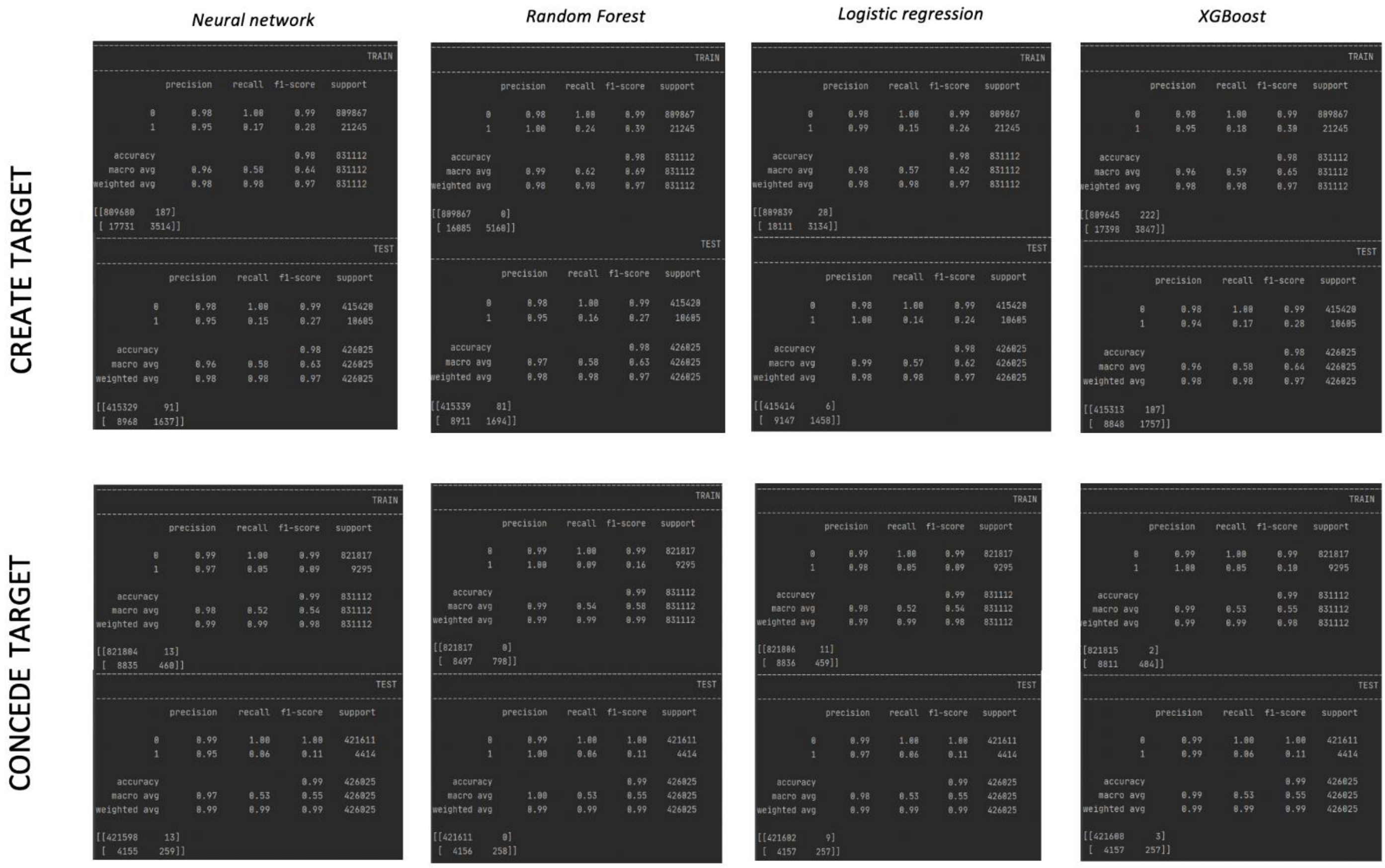


Figure 49 - Classification reports and confusion matrices

Figure 50 - Pitch coordinate system by Opta