FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Mining football players' behavioral profile: identifying candidate proxy features from event data

Luís Jorge Machado da Cunha Meireles

Mestrado em Engenharia e Ciência de Dados

Supervisor: João Pedro Carvalho Leal Mendes Moreira

September 30, 2022

Mining football players' behavioral profile: identifying candidate proxy features from event data

Luís Jorge Machado da Cunha Meireles

Mestrado em Engenharia e Ciência de Dados

September 30, 2022

Abstract

Within the last decade, the Football Analytics field has grown significantly. Today, several companies specialized in acquiring, treating, and selling football-related data that clubs can use to improve their player recruitment and development procedures. However, football performance is multidetermined and complex. Therefore, unearthing knowledge about the physical, technical, tactical, and, above all, psychological dispositions of football players from such data is not a trivial task.

Building upon the football analytics literature, especially the VAEP approach and Bransen et al. (2019) pioneering effort in modeling mental pressure, we propose our own method to model normative contextual pressure. Despite the obvious commonalities, we deviate from the above-mentioned approach by acknowledging the Transactional Model of Stress (Lazarus and Folkman, 1984). Hence, several important data preparation jobs were performed to generate features: estimating players' individual contributions, calculating win-draw-loss probabilities throughout football matches, and computing outcome probabilities for each team using the Monte Carlo method. While following the CRISP-DM methodology and throughout the reported work, we used several Machine Learning (supervised and unsupervised) methods to compute the features, estimate probabilities, linear relationships, and group data within business-relevant clusters.

In the end, we present a use case in which we group players into different profiles (each one determined by a cluster) regarding their ability to perform with varying levels of normative contextual pressure.

Keywords: Football Analytics, Data Mining, Psychological Profiling, VAEP approach.

Resumo

A área de Football Analytics cresceu significativamente na última década, pelo que, nos dias que correm existem várias empresas especializadas em adquirir, tratar e vender dados de jogos de futebol que os clubes podem usar para melhorar os seus procedimentos de recrutamento e desenvolvimento de jogadores. No entanto, o desempenho no futebol é multideterminado e complexo, e criar conhecimento sobre os fatores físicos, técnicos, táticos e, sobretudo, psicológicos embutidos em tais dados não é uma tarefa trivial.

Com base na literatura da área de Football Analytics, especialmente no que concerne à abordagem VAEP e ao esforço pioneiro na modelagem da pressão mental de Bransen et al. (2019), propomos o nosso próprio método para modelar a pressão contextual normativa. Apesar das semelhanças óbvias, desviamo-nos da abordagem acima mencionada ao sermos informados pelo Modelo Transacional do Stress e do Coping (Lazarus and Folkman, 1984). Nesse sentido, efectuaram-se algumas tarefas de preparação de dados importantes: estimação das contribuições individuais dos jogadores, cálculo das probabilidades de vitória-empate-derrota ao longo dos jogos, e computação das probabilidades para cada equipa em termos de resultados finais usando o método de Monte Carlo. Seguindo a metodologia CRISP-DM, usamos vários métodos de Machine Learning (supervisionado e não supervisionado) para calcular «features», estimar probabilidades, relações de causalidade e agrupar dados em clusters relevantes.

No final, apresentamos um exemplo de aplicação no qual agrupamos jogadores em diferentes perfis (cada um determinado por um cluster) tendo em conta os seus desempenhos em níveis variados de pressão contextual normativa.

Keywords: Football Analytics, Mineração de Dados, Perfis Psicológicos, Abordagem VAEP.

Acknowledgements

First, I would like to thank Professor Doutor João Mendes Moreira for all the guidance during the Master thesis elaboration and throughout the Master in Data Science and Engineering course. It was not an easy job, given my non-IT background!

I would also like to praise Tiago Mendes Neves for co-supervising me during the Master thesis. It was a huge opportunity to learn football analytics from someone with considerable experience in data wrangling tricky football data. Furthermore, all these heavy last months of work gave us what I see as a solid friendship.

Furthermore, I would like to thank FC Porto, especially the Board members Vítor Baía and Eng. Luís Gonçalves, for all the support and interest in Football Analytics. I hope this work can be the cornerstone for future developments regarding establishing a Football Analytics department or group work within the club.

A short but sincere compliment to all my colleagues from the first-ever Master in Data Science and Engineering class; we are a great team, fraught with good people.

Last but surely not least, I would like to thank my family for all the support. All of you have been incredible during these two very intense years, especially Ana, my beloved wife, and my mother, Rosalinda. I would also like to dedicate this work to Ana and our beautiful daughters, Rita and Margarida, for the love that gives meaning to everything.

Luís Meireles

“ Much of the best mathematical inspiration comes from experience and that it is hardly possible to believe in the existence of an absolute, immutable concept of mathematical rigor, dissociated from all human experience.”

John von Neumann

Contents

List of Figures

List of Tables

Abbreviations

API	Application Programming Interface
CRISP-DM	Cross-Industry Standard Process for Data Mining
EPV	Expected Possession Value
FIFA	Fédération Internationale de Football Association
GPS	Global Positioning System
LSTM	Long Short-Term Memory
SPADL	Soccer Player Action Description Language
SPI	Soccer Power Index
VAEP	Valuing Actions by Estimating Probabilities

Chapter 1

Introduction

> "I find it terrible when talents are rejected based on computer stats. Based on criteria at Ajax now I would have been rejected. When I was 15, I couldn't kick a ball 15 meters with my left and maybe 20 with my right. My qualities, technique and vision, are not detectable by a computer" (Johan Cruyff)

1.1 Context

Football evolved from a simple leisure activity into a highly specialized and demanding industry in about a century. As football players need to perform at a high level throughout long and demanding seasons, football clubs must strive as fully-fledged companies. Following FIFA's allowance for wireless sensors usage during official matches back in 2009, big clubs moved toward innovative and scientifically rooted solutions regarding player recruitment and development operations. For instance, using GPS systems for physical workload monitoring and event or tracking data (when not combined) for scouting and performance analysis became common practice.

From a business perspective, and considering the mutual dependence between evaluation and planning, the investment in cutting-edge knowledge-generation solutions is logical. Therefore, and speaking from a strictly sportive sense, clubs are interested in improving their knowledge and procedures regarding football players' recruitment and development. In this scenario, data mining emerged as a means to an end. Hence, during the last two decades, several companies providing data collection, distribution, and analytics services have matured, and research agendas regarding the game's physical, technical, and tactical components proliferated (see Tuyls et al. (2021)).

However, despite being addressed as critical as the physical, technical, and tactical components by numerous sources in the sports science literature (e.g., Meylan et al. (2010); Williams and Reilly (2000)), the football analytics community largely ignores football players' psychological performance. The only known exception is Bransen et al. (2019) approach to football players' performance across different levels of normative contextual pressure.

1.2 Goals

This academic essay seeks three purposes by tacking Bransen et al. (2019) approach as a starting point. The first is to clarify the challenges of evaluating football players' performance from a Sport Psychology perspective. The second is to identify the best proxies for a theoretically accurate judgment of football players' performance during individual and aggregated football matches (e.g., throughout a season) with varying levels of normative contextual pressure. Finally, the third is to conduct and illustrate a use case, in which a data-driven approach is followed to identify different profiles of psychological ability to deal with stressful competitive situations. Bearing the initial Johan Cruyff quotation, the ultimate aim of this academic essay is precisely to show that computers indeed can perform informative estimations of football players' performance. While not perfect, they still can add value to player recruitment and player development procedures.

1.3 Methodology

The methodology underlying the several analysis compiled in this thesis was the CRoss-Industry Standard Process for Data Mining (CRISP-DM), the most influential process from an industry point of view to conduct data analytics projects. It consists of six non-rigid sequential steps that shall be kept in the loop throughout a company's lifetime and will be described below.

1.3.1 Business Understanding

According to Shearer (2000) and Wirth and Hipp (2000), it is the most critical phase of any data mining project since it entails acquiring or developing the minimum required knowledge to conduct an analytic project in a given field. Without such understanding, framing the analytical work in an operationally accurate way can be pretty misleading. It comprehends the following sub-phases: determining business objectives, assessing situational conditions/requirements, establishing data-mining goals, and producing a project plan.

1.3.2 Data Understanding

Data understanding is when the analyst first familiarizes himself with the available data. It starts with an initial data collection, followed by their initial visualization and general descriptive analysis to detect data quality problems (e.g., missing information or outliers) and putative initial insights (Moreira et al., 2018).

Besides having the minimum required knowledge to conduct an applied data mining project, a correct diagnostic of data quality is also a significant step with direct implications for model building. While lacking the required operational/applied knowledge can lead the analyst towards false or irrelevant questions/hypotheses, an unsatisfactory data diagnosis can bias data preparation and model construction and, ultimately, make models useless. More often than not, a trade-off between data quality and model sophistication must be settled.

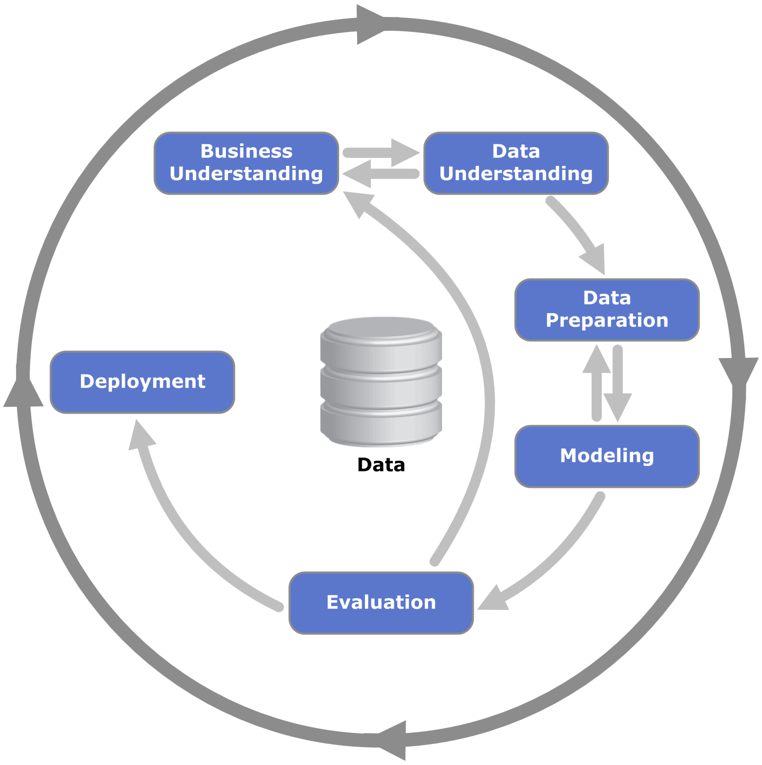

Figure 1.1: The CRoss-Industry Standard Process for Data-Mining (CRISP-DM) (from Shearer (2000))

1.3.3 Data Preparation

This step involves preparing the data set for the forthcoming modeling phase. During data preparation, analysts select the data they need and treat it in terms of outlier removal and missing data fulfillment, engineer features from the original attributes, integrate data from different data sources, and reformat data in case it is needed. While data understanding is more descriptive, data preparation is more operational-oriented with considerable data wrangling (Shearer, 2000).

Having the data properly clean and, to what's possible, noise-free (i.e., without defective samples and misread values) is mandatory for an accurate and smoother modeling phase. Furthermore, almost every data-mining problem requires the transformation of original attributes into new, more relevant features. For instance, in this thesis, we transformed data from game events into values attributed to individual football players' actions.

1.3.4 Modeling

Modeling is the phase where different modeling techniques are applied and tested to solve a particular problem or a set of issues regarding an initial data-mining goal. Here, several models can be tested against each other, which might entail additional data preparation tasks, given that each model has particular requirements/setting conditions (Moreira et al., 2018). Also, it is the phase where model-specific hyper-parameters are tuned.

1.3.5 Evaluation

The evaluation step has three sub-phases: evaluating the results, the review process, and determining the following steps (Shearer, 2000). As Moreira et al. (2018) pointed out, "solving the problem from the data analytics point of view is not the end of the process." Evaluating how well the model performs in an applied setting is necessary. Questions addressed herein are: is the model an" adequate"pill" for the initial business "pain"? What are the alternative solutions, or what can be done differently or tested in the future?

1.3.6 Deployment

Once a model is successfully evaluated, the deployment phase is reached. Herein, efforts are directed towards the accommodation of the obtained solution in the applied ecosystem (i.e., company processes). As the ecosystem changes, new relevant data-mining questions might arise, which, in turn, might lead to additional adjustments in each of the previous CRISP-DM cycle phases (Shearer, 2000; Wirth and Hipp, 2000).

1.4 Thesis Structure

The remainder of this thesis follows the structure detailed below.

- **Chapter 2** is a brief review of the literature regarding football analytics.
- **Chapter 3** while detailing Bransen et al. (2019)'s pioneering approach to model mental pressure, explains how the VAEP approach can be used to estimate individual performance under varying levels of normative contextual pressure.
- **Chapter 4** describes the method from the Business Understanding to the Modelling phase. It also presents the results for the estimation of features during Data Preparation.
- **Chapter 5** contains the exploratory data analysis work performed to model normative contextual pressure.
- **Chapter 6** details all the modeling procedure, and describes a use case in which players are grouped into different clusters of ability to perform under varying levels or normative contextual pressure.
- **Chapter 7** summarizes the work done and reviews the obtained results. It also entails a brief reflection on the limitations and putative future work.

Chapter 2

Data Mining for Football Analytics: A brief review

2.1 The nature of football data

Regarding football data analytics, two data types are the most typical: event stream data and optical tracking data (Decroos et al., 2019). The first is data describing the times and locations of with-ball discrete events (e.g., passes, shots, duels), while the last consists of the tracking records of players and the ball locations at a high-frequency rate using multi-camera systems.

Compared to event stream data, optical-tracking data has the advantage of providing synchronized information regarding each football player and the ball in a quasi-continuous fashion (typically 25 frames per second). Thus, robust estimations can be made using absolute (i.e., about a single player) and relational information (i.e., about the relationship between a player, the ball, and other players) (Rein and Memmert, 2016). Furthermore, while event data is biased towards offensive with-ball events, optical tracking data is not. Unfortunately, due to the high cost of optical tracking systems, tracking data is limited to wealthier teams.

As can be seen in Table 2.1, event data does not have the same accessibility issues as tracking data. It is generally cheaper, and data collection is spread across most football leagues worldwide. Nonetheless, its correct usage imposes several challenges, as reported by Decroos et al. (2019). First of all, the data-providing companies serve a multitude of clients (e.g., football clubs, the media, betting companies), and so their APIs are not optimized to ease data mining jobs; In second, the terminology used to characterize events is company-specific, which hinders data integration from different sources; The third challenge regards the inability of some event data platforms to adapt their platform architectures to new event stream formats, making querying operations quite ticklish; The fourth regards extra information snippets that some vendors provide that make automation very hard; And last but not least, most machine learning algorithms require fixed-length vectors, which is not granted from this kind of data.

Table 2.1: Event vs. Optical Tracking Data

	Event Data	**Optical Tracking Data**
Consists in	Discrete event logs	Quasi-continuous data streams
Pros	Cheaper Covers almost all leagues worldwide	Absolute and relational data
Cons	Not optimized for data mining jobs Data integration issues A fixed-length for vectors is not granted	Financially and computationally expensive

2.2 Data Mining for Football Analytics Using Optical Tracking Data

As the reader will see by checking the dates of the publications reported below, football data analytics using optical tracking data is only in its beginning. Nevertheless, some exciting research avenues blossomed in the last few years. Among other applications, the main areas of interest have been the following:

2.2.1 Identifying Team Formations and Game Phases

As with any team sport, the distribution of the players in the field has important implications given each player's functions and team tactical and strategical objectives. Thus, finding procedures that can classify the formations automatically when teams are attacking, defending, or transitioning from attacking to defending and vice-versa has been a goal for football data analysts.

Bialkowski et al. (2016) used several unsupervised learning algorithms (e.g., K-Means, Agglomerative Clustering) to compute players' deviations from the team prototypical formation. Likewise, Lucey et al. (2015) used unsupervised learning and conditional random fields to identify different defending team formations and game phases (e.g., defensive organization, transition, free kicks).

2.2.2 Calculating The Instantaneous Expected Outcome of Football Actions

To estimate or predict the impact of several football actions on goal-scoring is also an essential goal for data scientists enrolled in football analytics.

Fernández et al. (2021) used neural network architectures (both shallow and deep) to estimate the instantaneous expected outcome (i.e., Expected Possession Value (EPV) metric) of any soccer possession for passes, shots, and ball drives independently. Furthermore, the authors proposed a decomposed learning approach to generate visual interpretations for EPV (see Figure 2.1). Although requiring sophisticated encoding involving time series manipulation, the concept of EPV is simple: given a game state (s), predict which team will score at a future game time ($s + t$) regarding all the spatiotemporal information available (i.e., the location of the players and the ball, and the observed events).

Anzer and Bauer (2021), in their turn, introduced an expected goals (xG) model to estimate the quality of any given shot. Among the features used to train the model were: the coordinates of the shot location, the speed of the player attempting the shot, the number of defenders in the line of shot, the position of the goalkeeper, and the pressure on the player taking the shot. Using an ensemble algorithm (i.e., XG Boost), the authors' model calculated shots' goal probability better than other state-of-the-art approaches in the literature. In a somewhat related publication, Bauer and Anzer (2021) developed a model to detect counter-pressing situations. The authors framed the data mining problem as a classification task, using human-labeled data for training. After achieving a respectable classification performance, the authors estimated counter-pressing immediate direct consequences (i.e., performed shots).

Finally, Lucey et al. (2015) after dividing a football match into its corresponding game phases, created a model to estimate the goal probability of each shot, whether it happened after an offensive transaction or during the offensive organization.

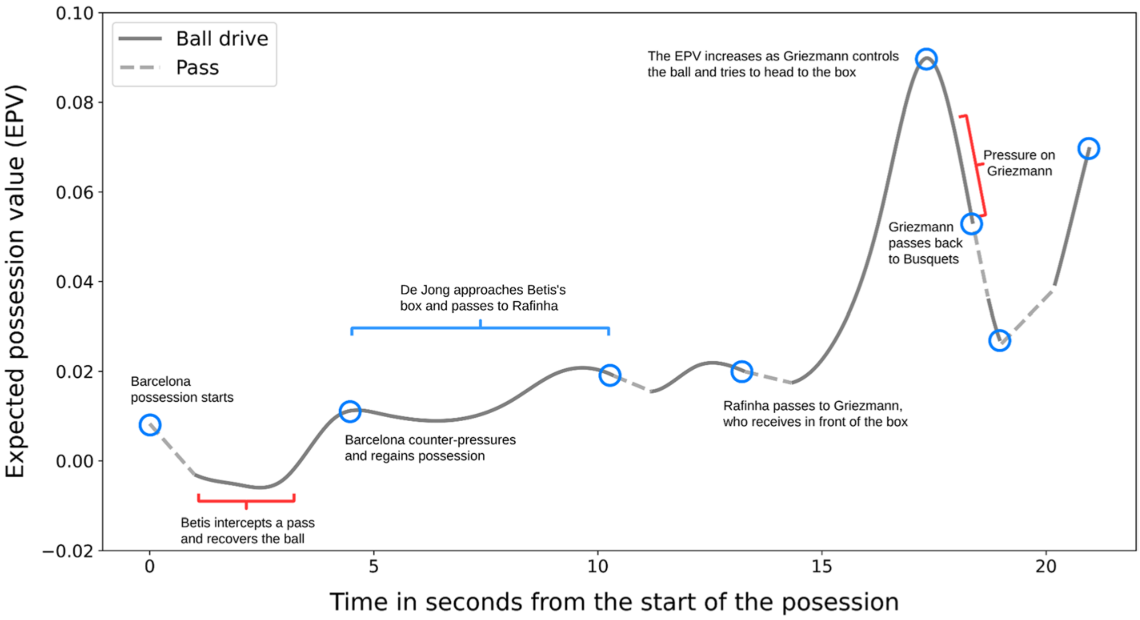

Figure 2.1: An example of play analyzed through EPV leans (from Fernández et al. (2021))

2.2.3 Pitch-Ownership and Pass Feasability Models

Pitch ownership models are proxy measures for the team's geographical pitch control. They evolved from Voronoi diagrams, which define *control* as a measure of distance to the closest player, reflecting the belief that the more effective a team is at controlling the space, the closer it is to scoring and not conceding goals. Pass feasibility (i.e., the ability to pass reach its target) relates to pitch control because pass feasibility depends on which team controls the space for a pass to meet its target.

Supported by Physics concepts and Markov processes, Spearman (2018), using both event and tracking data, crafted a measure to calculate the probability of a receiving player scoring in the next temporal window based only on the instantaneous game state. Brefeld et al. (2019) used a triplet codification approach with kernel density estimators to compute pitch control from positional data. Such a procedure improved the explainability level of traditional Voronoi diagrams (see Figure 2.2).

In two somewhat more computer-vision-oriented approaches, Arbués-Sangüesa et al. (2021) estimated pass feasibility in 2D maps given the positioning and body orientation of the receiving players using Gaussian functions, while Andrienko et al. (2017) hand-crafted a metric to calculate pressing. In its turn, Steiner et al. (2019) analyzed how different embedded passing options informed decision-making regarding the number of opponents outplayed from each pass.

Also, Fernandez and Bornn (2018) devised a sophisticated model to calculate the relative value of any pitch position given the ball's position and the relative importance of each zone through feed-forward neural networks. From this effort, two novel spatial value metrics emerged, one for occupation and the other for the generation of spaces. Spearman (2018) developed a ball control model that acknowledges the time to intercept from the defending adversary players and the time each attacking player needed to control the ball after receiving a pass. Goes et al. (2019) and Kempe et al. (2018) evaluate successful passes in terms of induced defensive disruptiveness for the adversary team (i.e., team spread).

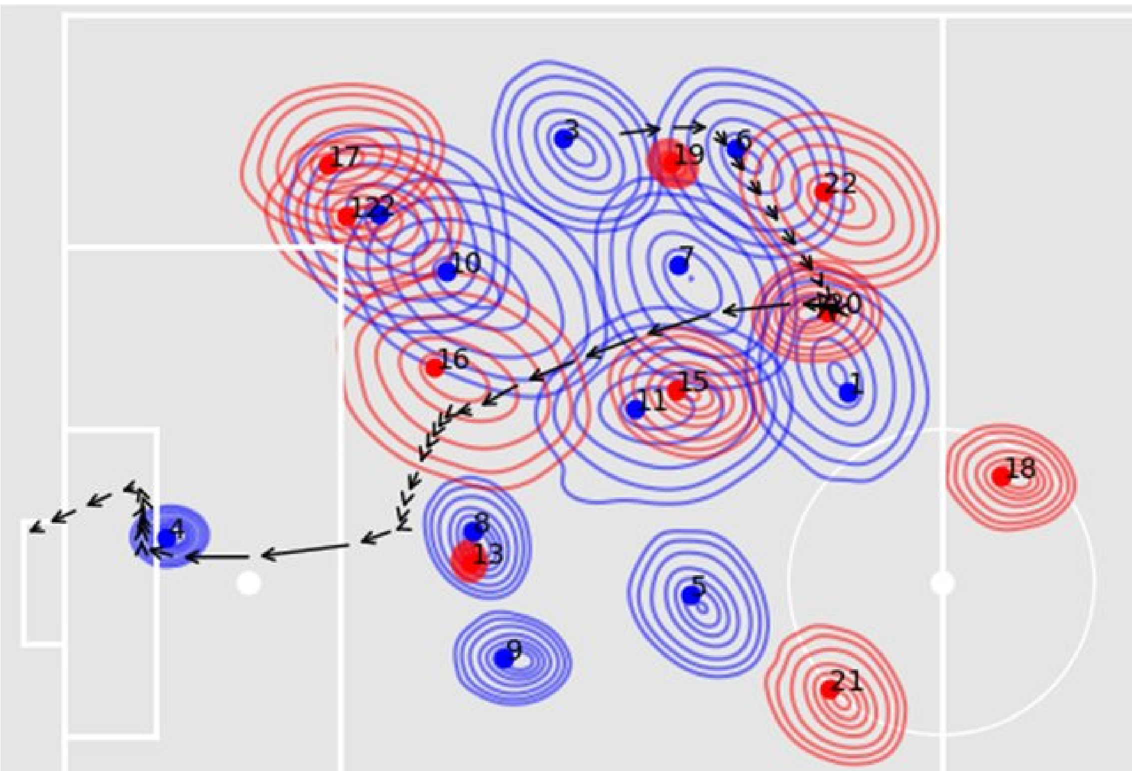

Figure 2.2: An example of a pitch control model (from Brefeld et al. (2019))

2.3 Data Mining for Football Analytics Using Event Data

2.3.1 Shooting and Passing Behavior

Brooks et al. (2016a,b) ranked football players taking into account made passes location and shot opportunities created by building a binary classification model to predict whether a completed pass would end in a shot or not. In a somewhat analogous fashion, Gyarmati and Stanojevic (2016) handcrafted "QPass," a metric capable of assigning a value to passes regarding their utility (moving forward in the field). By vectoring passes information for each player, Cho et al. (2022) used convolutional auto-encoders to estimate football players passing styles.

In a more sports science-grounded approach, Rein et al. (2017) estimated passing efficiency at a team level given distinct passing styles via Gaussian mixture models and the calculation of Voronoi diagrams. In their turn Szczepański and McHale (2016), in a more purely statistical approach, estimated the passing ability of soccer players by using generalized additive mixed effects to model passing difficulty and proxies for the opponent's defensive abilities among other factors from event data.

Duch et al. (2010) followed a different paradigm. Based on social network analysis, the authors developed a flow centrality metric to quantify the individual contribution of each player to team passing performance. In their turn, Link et al. (2016), in a more top-down approach, modeled a quantitative measure of dangerousity in offensive plays rooted in some sport science literature indicators (e.g., dominance, tackling rate).

2.3.2 Ranking Football Players Overall Contributions

Aiming at ranking football players, Pappalardo et al. (2019) devised and implemented a data-driven framework that offered a multi-dimensional and role-aware evaluation of soccer players' performance. Support Vector Classifiers were their choice for the classification tasks. In their turn, Liu et al. (2020), in a pioneering approach made under the umbrella of a deep reinforcement learning paradigm, evaluated all types of soccer actions from play-by-play data using a Markov Q function and two Long Short-Term Memory (LSTM) towers (i.e., one per team).

In a series of papers and conference papers, a research group associated with Scisports and to KU Leuven proposed the idea of concatenating in a single vector per player all the performed actions Decroos et al. (2019). From such a technique, several overall performance estimations were made using event data streams. Decroos and Davis (2019) used such approach to estimate soccer players' playing style, and several contributions were made to create player rankings (Decroos et al., 2019, 2020b,a,c). Such approach will be closely examined in the next section.

2.4 The Valuing Actions by Estimating Probabilities (VAEP) approach

The Valuing Actions by Estimating Probabilities (VAEP) approach framework revolutionized how event data was used for football data-mining operations (see Figure 2.3). As Decroos et al. (2019)

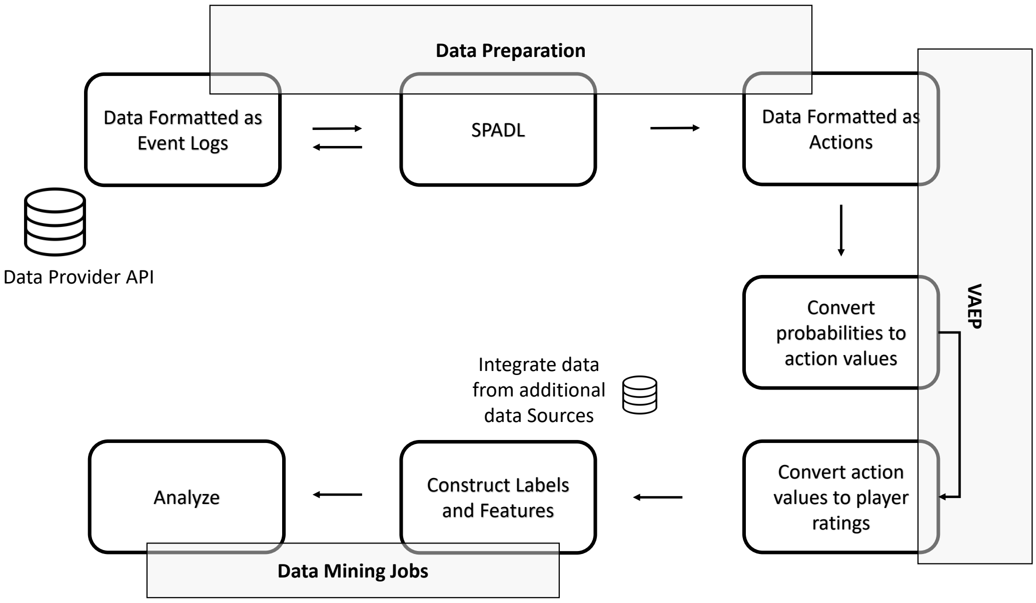

Figure 2.3: The VAEP approach outlined

reported, the traditional use of event data has several significant flaws, primarily because of two critical factors: 1) the low-scoring, and 2) the dynamic nature of football matches. Since goals are relatively rare events, it is not trivial to analyze how non-proximal events - that is, those that are temporally and geographically distant from goals - influence a team's probability of scoring or conceding a goal. For example, a fast sequence of passes in the middle may not lead to a goal in the next three seconds time window but can generate the necessary space to create a goal chance moments later.

By grouping into single feature vectors all the actions a player performs during a game, VAEP introduces a temporal dimension - at least, in a probabilistic sense. Specifically, by translating game events into subsequent actions, it is possible to train a machine learning model to assign a value to goal scoring and goal conceding probability for each action based on the difference in goal/conceding probability from two consecutive actions.

2.4.1 Translating Events into Actions

As reported at this chapter's beginning, mining event data entails several challenges. In sum, event data must be transformed to allow mining jobs, which sometimes entail massive data transformation procedures. It does not follow a uniform format between data providers, which hinders data integration. Moreover, the length of each event stream vector may not be fixed, which can impede the use of several machine learning algorithms.

Decroos et al. (2019), intending to uniformize the several event stream formats available in the market into a common language that eases data analysis, developed Soccer Player Action Description Language (SPADL). This language considers only events that describe actions performed by the players (e.g., passes, dribbles, shots), ignoring others such as the one that signals the end of the game. Indeed, by definition, "actions are a subset of events that require a player to act" (Decroos et al. (2019), p. 1852). Thus, operationally, a football match is a sequence of on-the-ball actions from $a_1, a_2, ... a_t$ where t is the total number of actions that took place at such a match. SPADL formats each action as a tuple of nine attributes:

StartTime: the action's start time,
EndTime: the action's end time,
StartLoc: the (x,y) location where the action started,
EndLoc: the (x,y) location where the action ended,
Player: the player who performed the action,
Team: the player's team,
ActionType: the type of the action (e.g., pass, shot, dribble),
BodyPart: the player's body part used for the action,
Result: the result of the action (e.g., success or fail).

2.4.2 Converting Probabilities into Actions Values

Both teams have two general objectives during a football match: scoring and not conceding goals. Hence, we may conclude that all individual on-the-ball actions intend to increase the odds of scoring and decrease the probability of conceding goals. Decroos et al. (2019) ingeniously deduced that a viable way to evaluate the effect of an individual action was by computing the difference between the actual and the preceding actions in goal scoring and goal conceding probabilities for own team. The assignment of a value (probability of scoring/probability of conceding) can be described formally as follows by Equation 2.1 (see Decroos et al. (2019)):

Given: an on-the-ball action a_i;

Do: Learn a function that assigns a value $V(a_i)$ to the action. Since every on-the-ball action alters the game state (s) consecutively, then:

$$V(a_i) = V(s_i) - V(s_{i-1}) \tag{2.1}$$

Each game state can be valued by the impact of such action on each team's probability of scoring and conceding a goal. If the action (the distance between s_{i-1} to s_i) positively affects team scoring odds, it shall be valued positively. Otherwise, it should be negatively valued. Then, as stated in Equation 2.2:

$$V(s_i) = P_{scores}(s_i) - P_{concedes}(s_{i-1}) \tag{2.2}$$

A further parameter (k) can be added to this equation, explicitly stating the number of succeeding actions considered when calculating the goal scoring/goal conceding probabilities for s_i. Therefore, we end with Equation 2.3:

$$V(s_i) = P_{scores}(^k s_i) - P_{concedes}(^k s_{i-1}) \tag{2.3}$$

2.4.3 Converting actions values to player ratings

Regarding player performance comparison, Decroos et al. (2019) proposed a method to rank players by their aggregated contribution to the team for several time granularities, which is formally described by Equation 2.4. The actions are normalized per 90 minutes of game time since more playing time represents more contributing opportunities. Formally:

$$rating(p) = \frac{90}{m} * \Sigma_{a_i \in A_p^T} V(a_i) \tag{2.4}$$

where:
A_p^T is the set of actions played during a given time frame T
$V(a_i)$ is the contribution value obtained by equation (2.3)
m represents the game time for a given player during T

Since in SPADL formatted data, there is information on the action type and the geographical coordinates in which each action takes place, additional analysis can be performed instead of simply summing all the actions. Player profiles can be traced based on such dimensions. For instance, it is possible to check where each player performs his most valuable actions in the field or to compare the relative contributions values for each type of action between several players (Decroos et al., 2019).

2.4.4 Constructing labels and features

Estimating the contribution of every individual action for the probability of a team scoring or conceding a goal in a given time frame k can be easily framed as two different classification problems. In the first case, that is, estimating $Pscores(S_i, x_i)$, a positive label (=1) is assigned every time a team scores a goal within k subsequent actions; otherwise, a negative label (=0) is set. The second classification problem, estimating $Pconcedes(S_i, x_i)$, follows the same approach (Decroos et al., 2019).

Decroos et al. (2019) proposed three classes of features that influence the probability of a team scoring or conceding a goal given a predetermined action lag (k).

- **Features embedded at SPADL**: Information explicitly retrieved from SPADL data representation (i.e., actions start and end locations, actions starting time, type of actions, and body part used).

- **Complex features**: Information derived from several variables within and across subsequent actions (i.e., distance and angle to the goal, distance covered during the action in both x and y directions, distance and elapsed time between consecutive actions, and change of possession).

- **Contextual features**: Information regarding the actual game state (i.e., the number of goals scored by the attacking team after a_i, the number of goals scored by the defending team after a_i, and the goal difference.

2.4.5 Analyzing Data

Once the scoring and conceding probabilities are estimated and each action accordingly valued, we can proceed to data mining jobs. Decroos et al. (2019, 2020c,a) and Bransen et al. (2019) provide some case studies that use VAEP to make sense of event data. For instance, Decroos et al. (2019) used VAEP to create player rankings, identify promising young players in minor leagues, and characterize football players' playing styles. On its hand, Bransen et al. (2019) used VAEP to make recommendations on player acquisition, player development, and tactical and strategical decision-making by analyzing how VAEP fluctuated across different levels of contextual pressure. The next chapter will perform a deeper analysis of the last referred work.

Chapter 3

Using VAEP to estimate the behavioral correlates of individual performance under varying normative contextual pressure conditions

3.1 Introduction

As stated in the introduction of this essay, although always referred to as one of the supporting pillars of sports performance, the psychological aspects are rarely objectively quantified. Nonetheless, both psychologists and the common sports audience acknowledge that performing in competitive settings (e.g., professional football) entails being able to cope with contextual pressure (Cruz, 1996; Gucciardi and Dimmock, 2008). By taking a nomothetic approach and assuming that situational pressure can be modeled, it is possible to construct a universal method or a general metric to estimate individual performance under varying levels of contextual stress.

While taking the VAEP framework to measure the value of each performed action (i.e., to compute the label or dependent values), Bransen et al. (2019) provide three types of applications or use cases through which such a metric could benefit football clubs:

- **1. Recruitment** - coping effectively with pressure is a very desirable skill that can be taken into account while recruiting new players;

- **2. Development** - the assessment of football players' behavioral profiles under situational pressure will reveal which players need to develop their coping skills further;

- **3. Tactical/strategical planning** - a measure of relative situational pressure based on actions would enable identifying which actions foster a team's probability of scoring under varying levels of contextual stress.

Furthermore, it would also inform decision-making regarding the composition of the starting lineups and substitutions according to the momentaneous situational pressure.

3.2 Modeling Normative Contextual Pressure in Football Matches

According to Bransen et al. (2019), two different types of factors affect contextual pressure levels: the pre-game factors and the in-game factors (see Figure 3.1).

- **Pre-game Factors** A varying level of *a priori* contextual pressure is associated with any football match. For instance, theoretically, a game involving two rival teams striving for the championship will be tenser than a game involving two middle-table teams when nothing (e.g., the danger of relegation, possibility of qualifying for the European competitions) is at stake. Bransen et al. (2019) identified team ambition (i.e., the a priori ambition a given team has regarding a competition), game importance (i.e., the value of each match for the team to meet their ambition), recent performance, and game context (e.g., home or away, rivalry, the match attendance) as determining factors pre-game contextual pressure.

- **In-game Factors** Virtually, whatever happens in a football match will impact the contextual pressure imposed on the two confronting teams. For instance, if a game is tied till close to the end, the game will probably become tenser in the remaining time since a goal can dramatically change the outcome of the match for each team. Conversely, if a team is winning by a substantial margin, the pressure will tend to decrease since a putative additional goal loses relative importance for the game outcome.

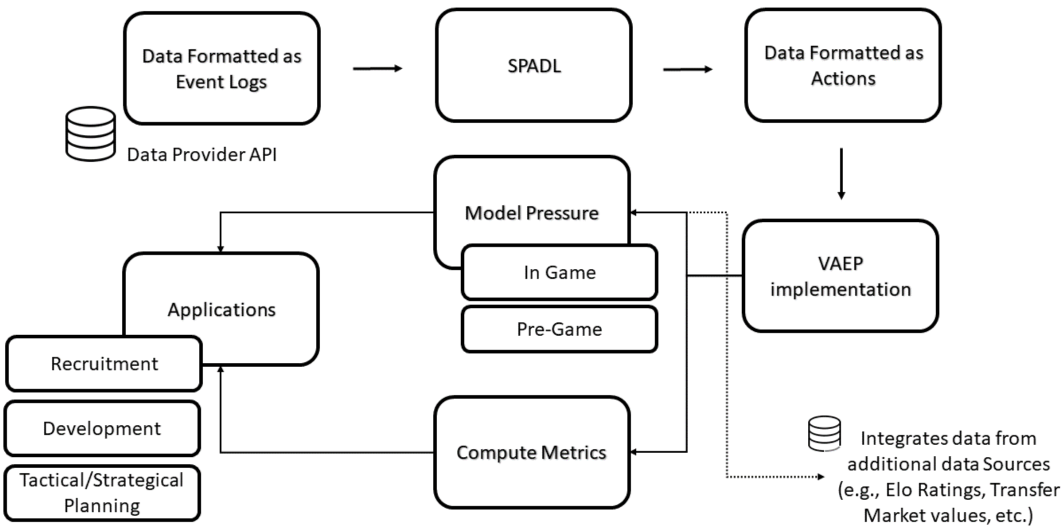

Figure 3.1: Bransen et al. (2019)'s approach to modeling situational pressure

Hence, contextual pressure can be conceived as the intersection, or the product, between pre-game and in-game contextual pressure. Formally, as described by Equation 3.1:

$$Pressure(g, x_t) = Pressure_{pre-game}(g) * Pressure_{in-game}(x_t) \tag{3.1}$$

where:
g stands for any given game
x_t stands for any game state

3.2.1 Pre-Game Contextual Pressure

Since there is no *a priori* labeling of each game's pressure level, Bransen et al. (2019) considered that a given set of games could be sorted by pressure level using pairwise rankings. Given the lack of ground truth, a panel of 19 soccer experts was consulted to build and evaluate a ranking classifier. The final model encompassed the abovementioned pre-game contextual pressure factors: several proxies for team ambition, game importance, and game context, and an aggregate measure of recent performance. The classifier was able to mimic soccer experts' inter-rater agreement.

3.2.2 In-Game Contextual Pressure

A fundamental property concerning any emotional state (anxiety or perceived pressure included) is dynamism. Emotions are not rigid, so they change as persons and circumstances interact and influence each other, altering the psychological meaning of situations (Lazarus, 2000; Lazarus and Folkman, 1984). Likewise, during a football match, it is accurate to suppose that pressure levels will vary over time as game situations change. Considering that goals have a much more significant impact on contextual pressure in close games than in very unbalanced matches (i.e., when a team is winning or losing for a large margin), Bransen et al. (2019) estimated the impact of goals on the expected match outcome by measuring the discrepancy in winning probability between the current game state and the two hypothetical game states consisting of each of the competing teams scoring a goal.

Instead of directly modeling the win-draw-loss probabilities, the authors estimated the number of goals a team would score to capture the uncertainty of win-draw-loss outcome in close matches. Specifically, two independent Binomial distributions regarding the goals expected to be scored by each team were obtained. Formally, the distributions are defined by Equations 3.2 and 3.3.

$$y_{>t,home} | \theta_{t,home} \sim B(T - t, \theta_{t,home}) \tag{3.2}$$

$$y_{>t,away} | \theta_{t,away} \sim B(T - t, \theta_{t,away}) \tag{3.3}$$

where:
θ represents each team's scoring intensity in the t^{th} time frame
T represents the actual time frame[1]

The scoring intensities derive from the following features: number of scored goals, goal difference, number of yellow and red cards, the difference in Elo-ratings of the teams, the average number of attacking passes in the previous 10 time frames, and the average percentage of duels won during the last 10 time frames.

3.3 Measuring Performance

The effects of perceived pressure on performance are disparate. For instance, two well-known consequences are attention disruption and excessive self-focus on motor skills (see Carver and Scheier (2012) and Baumeister (1984)). Therefore, to measure football players' performance under normative contextual pressure, Bransen et al. (2019) proposes several metrics that try to capture different aspects of performance: the total contribution for the team's probability of scoring and not conceding goals, the quality of decisions, and the quality of execution.

3.3.1 Total Contribution

For the calculation of the total contribution, Bransen et al. (2019) et al. followed the VAEP approach. Regarding all the on-the-ball actions, the idea is to measure how each action altered the team's likeliness of scoring and not conceding a goal in the near future. Each action is described by several factors, including its type, the body part used to perform it, and the location in the pitch.

3.3.2 Decision Rating

Decision-making and execution (or performance) are related but separate constructs (Endsley, 2017). For instance, a player can make a poor passing decision and yet execute the pass effectively. Conversely, a player can make an accurate and tactically relevant decision but fail in its execution. While calculating the decision rating, Bransen et al. (2019) is not interested in the result of a given action but in how each chosen action relates to other putative actions in terms of its tactical/strategical relevance measured by how much it alters its contribution to the team scoring likelihood. Hence, Decision Rating is defined by equation 3.4.

$$DR(s_i, a_i) = E[CR|s_i, a_i] - E[CR|s_i], \tag{3.4}$$

where:
a_i stands for a given action
s_i stands for a given game state
$E[CR|s_i, a_i]$ is the chosen action's expected contribution rating
$E[CR|s_i]$ is the expected contribution rating for other hypothetical actions

3.3.2.1 Expected contribution for the chosen action

Bransen et al. (2019) circumvent the problem of not having the teammates' precise locations by predicting the following action's contribution given the current game state recurring to historical data for each type of action. A binary classifier was used to train a model for every kind of action (e.g., pass, shot, take-on). Bearing in mind that every action can be successful or unsuccessful, the expected value of each its contribution rating is the weighted sum of the contribution of both outcomes is given by Equation 3.5:

$$E[CR|s_i,a_i] = P(o_i^+).CR(s_i,a_i,o_i^+) + P(o_i^-).CR(s_i,a_i,o_i^-), \tag{3.5}$$

where:
$P(o_i^+)$ is the probability that action a_i succeeds
$P(o_i^-)$ is the probability that action a_i fails

3.3.2.2 Expected Contribution for all the actions

Regarding estimating the expected contribution rating across the possible actions in a game state, the authors used historical observations of actions performed in similar games. The features: the current location of the ball, start and end locations of the previous two acts, the types of the last two actions, and the speed of the sequence, were used to train a Gradient Boosted Trees model to make the predictions.

3.3.3 Execution Rating

Regardless of the quality of the decision, the execution rating attempts to measure the quality of execution of each action. In that sense, this metric will reward successful efforts and punish those which not succeed. Therefore, mathematically, the execution rating is the difference between the observed outcome of an action and the predicted probability that it would be successful as defined in Equation 3.6:

$$ER(a_i,o_i) = [0_i^+] - P(o_i^+), \tag{3.6}$$

where:
$P[o_i^+]$ takes the value of one if o_i succeeds and is zero otherwise
$P(o_i^+)$ is given by the action success predictor referred to in the previous session

3.4 Measuring Performance Under Normative Contextual Pressure

The performance metrics and pressure model must be associated to analyze players' performance under varying levels of situational pressure. According to the pressure model, situations can be ranked by pressure level. Bransen et al. (2019) used the following cut-points to label each situation's pressure level:

- **High pressure**: above percentile 80
- **Normal pressure**: between percentile 20 and percentile 80

- **Low pressure**: below percentile 20

Then, the authors proposed aggregating the three previously calculated performance metrics for a given player or team to build insights regarding the applied cases referred to at the beginning of this chapter (i.e., player recruitment, development, and tactical/strategical planning). In the Figures 3.2, and 3.3, some use cases presented by the authors can be consulted. In Figure 3.2, we can see how Bransen et al. (2019) used VAEP contributions through varying levels of pressure to find an adequate substitute for a leaving player. In contrast, in Figure 3.3, the authors compare several players in terms of their performance under pressure.

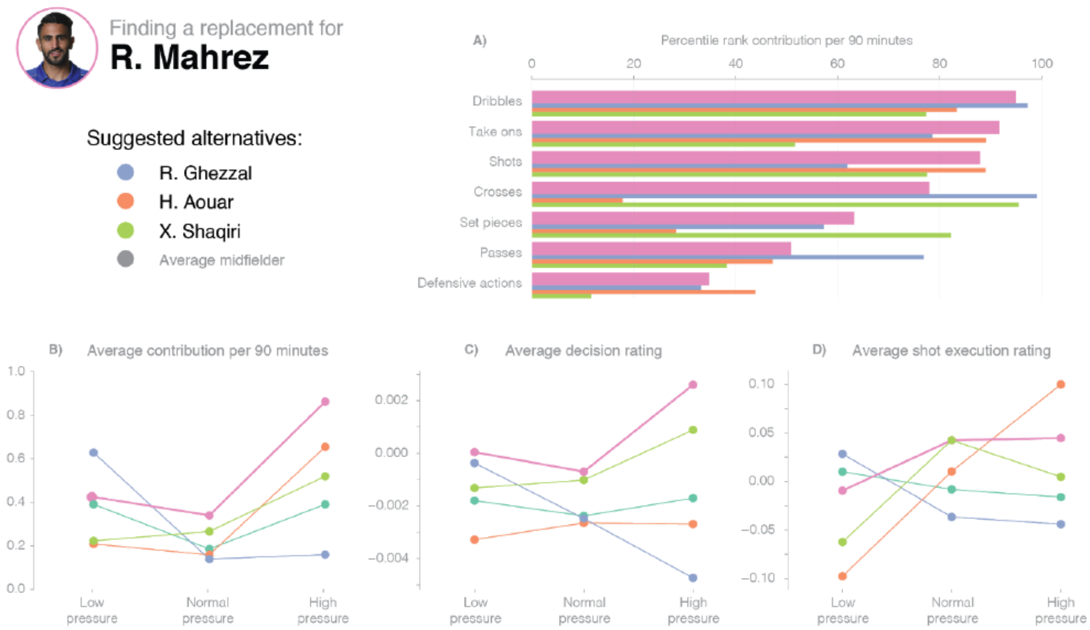

Figure 3.2: Use case: Recruiting a substitute for Riyad Mahrez (from Bransen et al. (2019)).

Figure 3.3: Use case: What are the players that perform best under high situational pressure (from Bransen et al. (2019)).

3.4.1 A critique based on the Psychology literature

Observation is a classical data-gathering approach in Sport Psychology. For instance, Smith et al. (1977) developed an observational system to assess coaches' behaviors, and Smith et al. (2015) validated an instrument to evaluate team motivational climate through observation. Regarding sports analytics, the observational power increment is a central contribution of recent technological advances. Nowadays, and compared to a decade ago, football players' behavior can be systematically annotated, tracked, and recorded in a finer-grained and rigorous fashion (e.g. Anzer and Bauer (2021); Tuyls et al. (2021)). When carefully handled, such data can provide valuable insights into football players' psychological performance estimation during competition.

However, when focusing on the psychological dimension of competitive performance estimation, "carefully handling" event and optical tracking data for football players' mental performance involves observing several inherent properties of human psychology and some necessary remarks regarding its evaluation.

The first idea to discuss here is the use of proxy measures to estimate the psychological performance of football players. Just as Physicists study black holes by looking at how they affect other stellar corpus, Psychologists deduce individual internal organization by looking at its behavioral manifestations (Houwer (2011)). They use behavior as a proxy for personality or subjective cognitive-affective processing system (Mischel and Shoda (1995)). Even in psychotherapy, the psychological evaluation process is proxy-oriented. The therapist is inferring when listening to the client's thoughts and feelings rather than measuring a physical property of the shared content. In other words, obtaining a global picture of an individual's mental life is, in practical terms, an approximation exercise guided by theory - that is, from the exhibited discourse and behavior during a session or by answering the items of a given scale, the psychologist estimates its underlying structural properties (e.g., Lingiardi and McWilliams (2017); Mahoney and Craine (1991)). The same reasoning is valid for the Sport Psychologist analyzing a football player's performance during a training session or a football match.

Nevertheless, and according to Houwer (2011)'s words, "using behavioral effects as a proxy for mental constructs violates the general scientific principle that the *explanandum* needs to be kept separate from the *explanans*" (p.203) and must be performed wisely. Bransen et al. (2019) intelligently avoid making a dangerous incursion into the realm of the subjective internal organization by assuming a functional approach, in which mental pressure stands as a contextual rather than an individual feature. The authors proposed a metric to compare football players' behavior as some situational features (i.e., rivalry, deviation from team goals, game importance, game's current state, recent performance) change without elaborating on what internal factors may explain such variation.

Aggregating "similar" situations in terms of contextual pressure provide some control to conduct *a posteriori* psychological studies regarding how football players react to such conditions. For instance, after creating groups for the best, average, and worst performers under high-pressure situations, studies can be conducted to evaluate their cognitive and affective profiles, developmental histories, and executive functions. Besides, the expert performance approach (see Ericsson et al. (1993, 2018)) follows the same differential method to study the development of expertise.

Another positive and theoretically accurate aspect of Bransen et al. (2019) proposal regards their "nonjudgmental" conception of situational pressure. If following a colloquial and stereotyped notion, the authors could have started from the idea that the effect of contextual pressure on performance would always be detrimental; however, as the authors found out, that is not the case. As Sport Psychology literature acknowledges, some athletes' performance improves as contextual pressure increases (Cruz (1996); Raglin (1992)). As made explicit by the Transactional Model of Stress and Coping (Lazarus and Folkman (1984)), anxiety experience is influenced by a critical mediator: subjective cognitive appraisals. Briefly, this theory suggests that contextual stress is not independent of how important the individual judges the situation for his own goals and well-being (i.e., primary cognitive appraisals), nor the perception concerning his ability to cope with the problem it embeds (i.e., secondary cognitive assessments). From this, two football players can experience similar events in opposite ways either by attributing divergent personal significance to the match (in terms of their own goals), having different perceptions regarding their ability to cope with the competition stressors, or a mix of both (Lazarus (2000)).

Finally, a somewhat obvious notion regarding the player performance measurement, but significant from a Sport Psychology viewpoint, is the distinction between the quality of decision and execution (see Chomsky (2014); Marr (2010)). Once again, the authors successfully addressed decision-making and action execution as separate constructs. A decision is in the realm of intention, depending not only on contextual information but also on factors beyond the specific situation (e.g., level of expertise, team strategy), and is not directly observable. On the contrary, individual actions are in the realm of execution, depending on technical ability and other contextual aspects (e.g., the atmospheric conditions, the adversary team's capability of conditioning the performed action), and are directly observable.

In sum, we found Bransen et al. (2019) contribution effective and promising, showing the path for further developments in data-driven solutions. Regarding the correct usage of behavioral data for psychological performance estimation, it: 1. assumes a functional approach, according to which the causes (i.e., independent variables) are within the domain of the situation, with individual behavior being the observable consequence (i.e., dependent variables). 2. it acknowledges the necessity of control (i.e., aggregation) regarding the situational conditions to compare football players' behavior. In other words, distinguishing between different situational conditions (even artificially) is a necessary preliminary step for computing meaningful individual cross-situational profiles. 3. it deals with performance, not the individual internal organization. 4. regarding decision-making, it is crucial to separately acknowledge the domains of intention (i.e., perception and cognition) and execution (i.e., performance).

The only critique we make to Bransen et al. (2019) approach is that when modeling situational pressure, the authors did not acknowledge the role of secondary cognitive appraisals (i.e., the judged ability to deal with the challenges imposed by stressful situations). Given the well-documented positive relationship between self-confidence and performance, Bransen's modeling of situational pressure seems unbalanced towards the more primary cognitive appraisals (i.e., the level of stress "directly" imposed by the situation - i.e., the importance and difficulty of each game). Therefore, when modeling normative contextual pressure (see Chapter 4), we introduced a feature that tries to capture the state of self-confidence of football players for a considered time frame.

Chapter 4

Methods and Results for Data Preparation

As stated in Chapter 1 , we selected CRISP-DM as a methodological approach. Therefore, this chapter follows the CRISP-DM structure.

4.1 Business Understanding

4.1.1 Business Objective

The main business objective of this thesis is to create a metric or a group of metrics that can capture football players' psychological performance during football matches. Such a metric or aggregate set of metrics could add value to a football club player's recruitment and development processes. Besides, it could also deliver relevant information for coaches' tactical and strategical guidance.

Establishing stringent business criteria for this goal is not straightforward due to the lack of ground truth. Although inspired by the only known contribution to measuring football players' performance under normative contextual pressure (Bransen et al., 2019), this work deviates from such work. Precisely in what concerns the features and algorithms selected to calculate individual contributions (i.e., VAEP value), modeling contextual pressure, and finally, in the computation of the performance metrics. Therefore, an objective direct comparison between the two approaches is infeasible in most cases. Nevertheless, by providing football clubs with a reliable quantitative indicator of individual performance under varying levels of situational competitive pressure, we are directly adding value to the business process listed above (i.e., player recruitment, player development, tactical/strategical guidance), which can be considered as a *de facto* business criteria.

4.1.2 Data Mining Goals

We must reach several data mining goals to accomplish the business objective mentioned in the last section. Specifically, to model normative contextual pressure, and create metrics that capture divergent aspects of performance (i.e., decision, execution).

Again, due to the lack of ground truth, there is not possible to use any standard metric to evaluate the models and metrics created. Validating contextual pressure would imply having data regarding football players' primary cognitive appraisals of threat, which we lack. The validation of our models would require the football players' to answer one or several measures of the perceived threat and perceived challenge for each match. Then, we could compare whether the developed metrics predict or correctly classify football players' scores. Ideally, it is something that we shall do in the future.

4.1.3 Tools and Situation Assessment

For this project, we have used the following tools:

- Python programming language;
- Numpy, Pandas, Matplotlib, Seaborn, Sklearn, Socceraction python libraries.

We used free licensed data from:

- PublicWyscoutLoader (made available by the Socceraction library);
- Fivethirtyeight.com.

4.1.4 Evaluation Metrics

Regarding model evaluation, we used several evaluation metrics. All these metrics are standard throughout disparate literature within the data science community.

4.1.4.1 Correlation of Determination

The correlation of determination (denoted R squared) corresponds to the proportion of the variation of the dependent variable that is predictable from the independent label. It is calculated as the ratio between the sum of the residuals squared and the sum of the distance the data is away from the mean also squared, providing a measure of "goodness of fit," that is: the distance between the estimated and absolute values as represented in Equation 4.1.

$$R^2 = 1 - \frac{sum\ squared\ regression\ (SSR)}{total\ sum\ of\ squares(SST)} \tag{4.1}$$

4.1.4.2 Mean Absolute Error

Considering that an absolute error is the difference between a measured and a 'true' value, we can say that the mean absolute error (MAE) averages all absolute errors. It is calculated as the sum of absolute errors divided by the sample size as stated in Equation 4.2.

$$MAE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.2}$$

4.1.4.3 Median Absolute Error

Instead of the mean, the median absolute error takes the median of all absolute differences between the target and the prediction. It is particularly interesting because, compared with MAE, it is more robust to outliers. As stated by Equation 4.3:

$$MedAE(y,\hat{y}) = median(|y_1 - \hat{y}_1|, \ldots, |y_n - \hat{y}_n|) \tag{4.3}$$

4.1.4.4 Accuracy

The accuracy is a standard metric in classification problems, which is the simple ratio between the number of correct predictions and the total number of predictions as stated in Equation 4.4.

$$Accuracy = \frac{correct\ predictions}{number\ of\ predictions} \tag{4.4}$$

4.2 Data Understanding

4.2.1 Data Collection

The Public Wyscout dataset is a public release of event data, compiled by Wyscout, that contains the events from all matches of the 2017/18 season of the top-5 European leagues (La Liga, Serie A, Bundesliga, Premier League, Ligue 1), the FIFA World Cup 2018, and UEFA Euro Cup 2016 (see Pappalardo et al. (2019) for a detailed description).

In his turn, the Fivethirtyeight.com dataset contains several complex statistics regarding football matches' result prediction like "Soccer Power Index," "Game Importance," and "Expected Goals."

4.2.2 Evaluation of the Collected Data

Regarding the general quality of the collected data, we have no significant issues to report (see Table 4.1). In the Public WyScout data set, notwithstanding the decent documentation of data, it was not easy to understand how Wyscout coded some important information snippets (e.g., yellow and red cards given). Concerning the timeliness of data, the data was not the most recent (i.e., 2017-2018 season), with extra features being available since that time (e.g., x,y coordinates of all players during each action). We believe such features would have improved the model of normative contextual pressure and the subsequent analysis compiled in this essay.

For the FiveThrityEight data set, we have to report the presence of missing data regarding the feature "importance." However, the magnitude of omitted values was not very expressive (i.e., 2,7%) and corresponded to the games of the same game week of the Italian Calcio A 2017-2018 season.

Table 4.1: Data quality for the Wyscout and FiveThirtyEight data sets

DataSet	**WyScout**	**FiveThirtyEight**
Accuracy	Data was accurate for each attribute	=
Interpretability	Available metadata (however documentation was unclear for some events)	Available metadata
Completeness	No missing attributes, nor missing data	2,7% of missing data for the "Importance" attribute
Consistency	Data followed a well-documented standard	=
Believability	Credible, and validated through peer-reviewed procedures	Credible
Timeliness	Data was somewhat old, newer datasets are available with relevant features; however, not for free.	Data was available for the relevant timeline.

4.2.3 Used Data

4.2.3.1 Public WyScout data set

In the Public WyScout data set, events are defined or characterized by twelve attributes. Amongst them are the team and player identification, the event type, and spatiotemporal tags (i.e., x and y coordinates and the time in which the event occurred in milliseconds) for each event (a sample of the data set is available in B). However, to calculate the individual contribution of each player for the team's probability of scoring and not conceding goals, we had to transform the data according to the SPADL language (see Chapter 2, section 2.4.1).

Shortly, SPADL distinguishes between player non-attributable occurrences (i.e., the event describing the start of the match) and actions performed by the players (e.g., a pass, a take-on), sorting data into the sequence of actions that happened in a game according to a tuple of relevant attributes. Therefore, we ended up with a transformed data set with labels regarding the identification of the player who performs a given action (e.g., game id, team id, player id) and the features of each completed action (e.g., start x location, start y location, end x location, end y location, time in seconds). A sample of the data set is available in C.

4.2.3.2 FiveThirtyEight data set

While Wyscout specialized in acquiring, organizing, and selling event data streams, the FiveThirtyEight website provides elements for forecasting soccer results. Within the context of this thesis, we used such data as a proxy to measure each team's strength (i.e., SPI rating) and the relative importance of each game in terms of the ending classification.

In this data set, two types of features were available. On the one hand, it had labels identifying each instance that corresponded to a given game (e.g., season, date, league, team1, team2, the game score), and, on the other, complex statistics regarding win, draw, and loss probabilities, the soccer power index (SPI), and the importance for each team.

The SPI rating estimates a team's overall strength by computing and combining an offensive and a defensive rating for each team. While the offensive rating corresponds to the number of goals a team is expected to score against an average team in a neutral field, the defensive rating consists of the number of goals a team is likely to concede in the same circumstances. From these ratings, it is possible to calculate the percentage of available points the team would be expected to take if the two teams played the same game in a loop. Given the SPI rating of two teams, the probable result of a match can be projected according to an Elo-model fashioned way (i.e., by direct pairwise comparison). Likewise, we can make simulations (e.g., using the Monte Carlo method) to calculate the probability of any competing team becoming a champion, qualifying for the European competitions, or being relegated. Before a season begins, a team's SPI ratings are calculated as defined in Equation 4.5:

$$SPI_{Pre-season} = \frac{2}{3} * SPI_{Previous\ Season} + \frac{1}{3} * SPI_{Market\ Value} \tag{4.5}$$

However, as the season unfolds, the ratings are adjusted after every match based team's performances and the quality of the opponents.

Importance measures how much the match's outcome will affect the team's statistical outlook on the season. FiveThirtyEight calculates the importance of a game by generating probabilities for each factor conditional on winning (or losing) the match and then finding the difference between those two possible numbers. The factor with the maximum range of the difference is taken, and the result is normalized between 0 and 100. Formally, the importance of a game for a team is given by Equation 4.6.

$$Importance = max(P(O|win) - P(O|lose)), O \in [Champion,\ UCL,\ UEL,\ Relegated] \tag{4.6}$$

4.2.4 Data Exploration

By exploring the Public WyScout data set, a first conclusion became clear. Due to the relatively small amount of data (only one season per competition plus two international competitions), we would need to use data from different leagues to train machine learning models while making predictions for a given target league. Although biasing the model by making it less specifical to the target league (e.g., different leagues may observe different distributions in terms of performed actions), which is not desirable, that was the only viable solution to provide the model with an adequate number of training instances. Within the available literature that used machine learning models to estimate or rank the football player's performance, the magnitude of training data was more extensive than ours. Nonetheless, the authors followed the same approach (see Bransen et al. (2019); Pappalardo et al. (2019)).

4.2.5 Data Preparation

To estimate the players' contributions to their respective teams' scoring and not-conceding goals (i.e., VAEP scores) and to model normative contextual pressure, we had to integrate data from several sources and perform several preprocessing jobs. For clarity, we organized this section according to such procedures, which are depicted in the Figure 4.1.

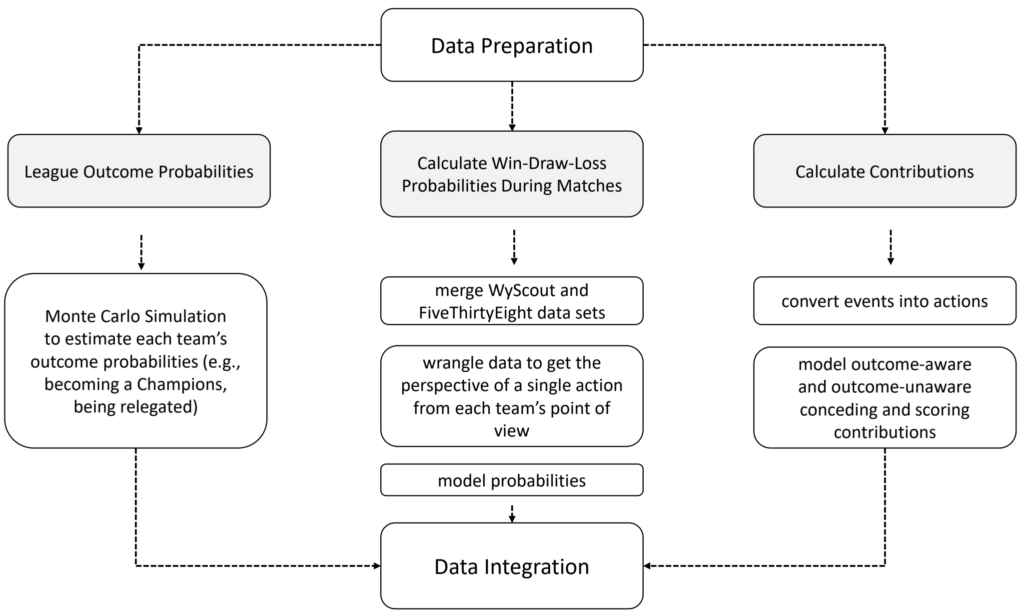

Figure 4.1: Scheme of jobs for data preparation

4.2.5.1 Calculate Contributions

We had to follow several steps to calculate how each action contributed to the probability of a team scoring and not conceding goals (i.e., VAEP scores). First, we loaded the game events from the

Public WyScout API using the 'socceraction' library. We then used the 'SPADL' module from the same library to convert events into actions. After data ingestion and formatting, we concatenated the necessary features and labels in the same data frame to implement machine learning algorithms regarding the estimation of such probabilities.

Although we aimed to adopt the original set of features and labels used by Bransen et al. (2019) and Decroos et al. (2019), such VAEP implementation follows a "brute force" approach, which forced us to adapt our method due to computational restrictions. Besides the actual actions, the authors considered complex spatiotemporal relationships between actions for the previous ten instances, exponentially increasing the complexity and the amount of features. Consequently, we had to limit the number of such features to the last three actions. For the labels, we used the predefined label (k=3) to make the predictions. Respectively, if a given action is one of the previous three before a goal, it is assigned to one class (value = 1). Otherwise, it is allocated to the other (value = 0).

Finally, we used the Random Forest Classifier to estimate the contributions of each action to the probability of scoring and conceding goals. Herein, we calculated two scores: an execution VAEP score and a decision VAEP score. Specifically, the execution VAEP score considers the result of each action (i.e., success or failure) as a feature to train the model, while the decision VAEP score does not. We have made such a distinction because we hypothesized that while the first would be a better proxy for execution quality, the second would capture aspects more related to the decision quality, which is a relevant distinction when analyzing football players' performance. For instance, a player may make a decision that would contribute much to his team's probability of scoring a goal but execute poorly. Conversely, he can perform a highly-skilled pass that contributes very little to his team's likelihood of scoring a goal in the near future.

Regarding the parameterization of the Random Forest Classifier, we used 100 trees as the number of estimators parameter and set the minimum number of samples required to split an internal node to 50. We do not carry any optimization procedure (e.g., random search, grid search) because we knew beforehand that those were the best parameters to tune the model (Mendes-Neves et al., 2022).

Table 4.2: Classification Model Results for VAEP scores estimation

		Train			Test		
		R2	**MAE**	**MedAE**	**R2**	**MAE**	**MedAE**
execution	**scoring**	0.339	0.023	0.008	0.144	0.026	0.009
	conceding	0.224	0.008	0.002	0.030	0.009	0.002
decision	**scoring**	0.256	0.026	0.008	0.040	0.029	0.010
	conceding	0.214	0.009	0.002	0.021	0.009	0.002

The results of the classification model can be consulted in Table 4.2. We can observe that the R2 score decreased from the training to the testing condition, which was expectable, and also that these values are low. In ordinary situations, such low scores would mean that the model cannot explain (or you only could explain very little) the relationship between the variance of the features and the target variables. However, we may argue, like in Mendes-Neves et al. (2022), that the R2 score is probably not the best metric to evaluate probability models regarding soccer data. Concretely, since goals are very infrequent events in soccer with such a significant influence on the calculation of the VAEP scores, and the R2 score is much sensitive to the variance, the MAE and MedAE errors may be better estimators. Looking at these values, we can see that they are pretty low and vary very little from the training to the testing conditions.

4.2.5.2 Calculate Win-Draw-Loss probabilities During Matches

To accurately model normative contextual pressure, besides the pressure inherent to each game, which regards the amount of pressure embedded within each match's importance and the opponent's team's relative strength, we also need to look at how pressure mounted or lowered as the game situations unfold (e.g., goals are scored, yellow and red cards are given)(Bransen et al., 2019).

One way of looking at this kind of pressure is by estimating how win-draw-loss probabilities evolved as the game progresses. Regarding such an estimation, we took several steps.

First, we merged data from the WyScout and the FiveThirtyEight data sets to later estipulate the features to calculate the above-cited probabilities. While in the last, we had information regarding the importance and strength of each team; in the first, we had information regarding the events that could alter the level of contextual pressure during the game.

Later, we had to transform the data to obtain each team's perspective on the ongoing result to calculate the win-draw-loss probabilities for each team. In the original event format data, each row corresponds to an event assigned to a team. However, to calculate the win-draw-loss probabilities for both teams, we must consider that changes in the result have opposing meanings for each team. For instance, a goal scored means getting closer to winning the game for the scoring team and the opposite for the team that suffers the goal.

After performing that transformation, we calculated the probability of winning, drawing, or losing the game from the point of view of the team performing each event. For that, we trained a multiclass classification model with each team's SPI, the current time, the current score, and visiting/visited condition as features, and the result (win, draw, or loss) as the label.

We tested several algorithms' performance with Python's sklearn standard parameterization for this machine learning job, using 10 Fold cross validation. In Table 4.3, we compare their performances in terms of average accuracy.

Table 4.3: Classification model results for win-draw-loss probabilities estimation

	Training Results	
	Avg. Accuracy	**Standard Deviation**
Random Forest	0.865	0.0538
XG Boost	0.639	0.0585
Logistic Regression	0.594	0.0492
Kneighbors Classifier	0.493	0.0537

We choose the Random Forest Classifier to estimate the probabilities for the testing situation from the observed results. It evidenced a better average accuracy score and similar standard deviation compared to the other algorithms.

4.2.5.3 League Outcome Probabilities

According to James (1980), “a Monte Carlo technique is any technique making use of random numbers to solve a problem,” where a random sequence of numbers is a finite consecutive set of numbers generated by chance. In other words, it can be said that a Monte Carlo estimation of F is a function of the sequence of numbers $(i \ldots +\infty)$ used in such an operation.

The use of random sequences of numbers for probability estimation has a long tradition in estimating purely deterministic yet complex probability distributions. As referred by Kroese et al. (2014) Monte Carlo techniques operate by “repeating the experiment many times to obtain many quantities of interest using the Law of Large Numbers and other statistical inference methods.” In short, the accuracy of the prediction of a target value (e.g., a probability) increases as a function of the performed simulations, as defined in Equation 4.7, where x is the number of simulations.

$$lim_{x \to +\infty} P_{MonteCarlo} \equiv P_{True} \tag{4.7}$$

We conducted Monte Carlo simulations to calculate each team’s outcome probabilities (i.e., the probability of becoming Champions, qualifying for the Champions League, qualifying for the Europa League, or becoming relegated) for the target season. We generated simulated league tables for each round from the FiveThirtyEight data set’s winning, drawing, and losing probabilities.

Chapter 5

Exploratory Data Analysis

5.1 Contribution Ratings

Regarding the contribution ratings (i.e., VAEP execution and VAEP decision scores), we performed the following analyses: comparing VAEP execution and decision distributions for some of the most relevant actions for VAEP (i.e., shots, passes, and take-ons) and exploring how VAEP changed over time and space.

As seen from the Figures from 5.1 to 5.3, we have distributions significantly skewed to the left for all the considered actions, which means that we generally have lower than highly valued scores. Therefore, a first impression we can get from the quality of the computation of the VAEP scores is that it accords with the nature of football as a low-scoring game. Given that we have defined each action contribution as the probability of that action leading or avoiding a goal in the upcoming three actions, distributions like those obtained were expectable.

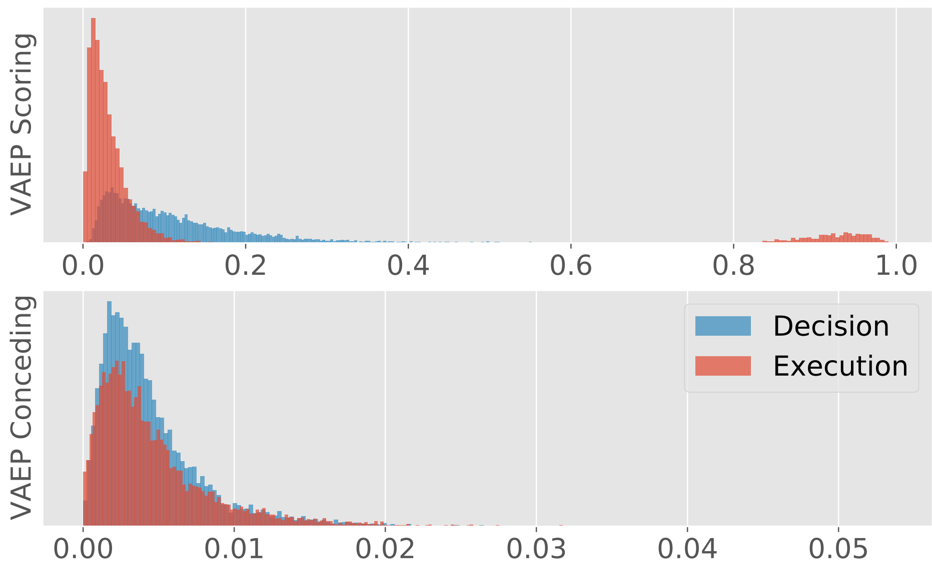

Figure 5.1: VAEP decision vs. execution distributions for shots (notice the different scales)

Figure 5.2: VAEP decision vs. execution distributions for passes (notice the different scales)

Furthermore, we can observe that the information regarding the outcome of each action is particularly relevant for the VAEP scoring condition. When such information is used to calculate the VAEP score (i.e., VAEP scoring execution), the distribution tends to be more skewed in the execution condition than in the decision condition, which is also expectable. Therefore, we can say that the VAEP scoring execution score is more stringent than the VAEP scoring decision score because fewer actions actually lead to goals than those that can potentially and eventually do it.

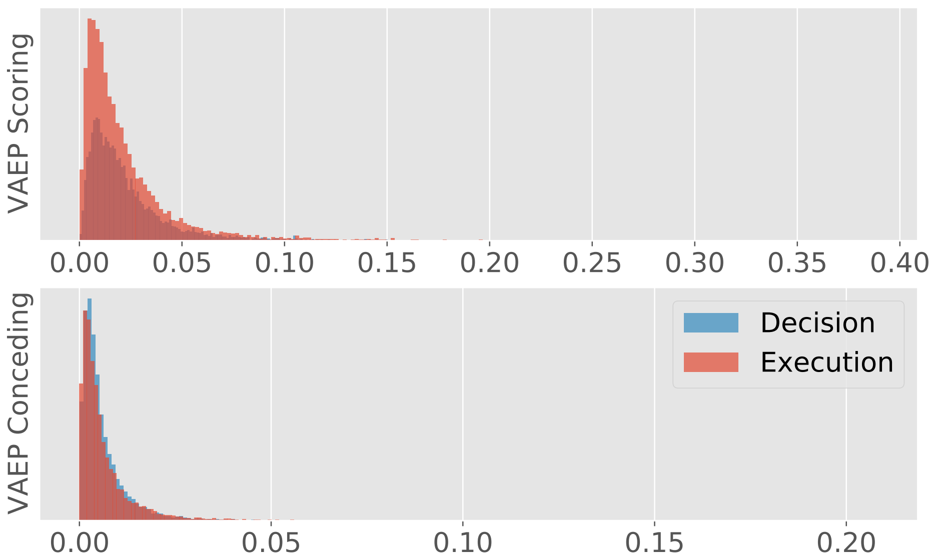

Figure 5.3: VAEP decision vs. execution distributions for take-ons (notice the different scales)

At this point, it is also relevant to determine whether the different VAEP scores correlate and interpret the meaning of the observed correlations if that is the case. As observed in Figure 5.4

and as expected, the two conceding (decision vs. execution) conditions are related, and the same happens for the two scoring conditions (decision vs. execution). However, the magnitude of the correlation observed is not the same. While in the former case, we observe a high correlation between the decision and execution conditions, we observe only a moderate correlation value in the latter.

From this, we may presume that information about the success/failure of a given action is likely more relevant to offensive than defensive moments. Given the magnitude of the correlation between the VAEP conceding decision and the VAEP conceding execution, we can even question whether or not it makes sense to proceed with the distinction between the two. While in theoretical terms, the distinction between the execution and decision condition makes absolute sense (Endsley (2017)), from the practical point of view and for the considered data, we were probably not dumping much information if we discarded one of them.

Figure 5.4: Correlation matrix for the different VAEP scores

Finally, it is also relevant to see how contribution ratings (i.e., the VAEP scores) vary according to time and space. Regarding time, and as observed in Figure 5.5, the VAEP scores tend to be higher in the second half. It makes sense that games get riskier (i.e., more chances of scoring and conceding goals) as they approach their end. On the one hand, fatigue increases the chances of errors, while on the other, there is less time to improve the game outcome for the team losing the match. Therefore, losing teams tend to choose riskier strategies as the game approaches the end. While increasing their chances of scoring, they may also increase their chance of conceding if the opposing team successfully manages to take advantage of the assumed risk (e.g., by effectively counter-attacking). Indeed, the odds of the gambling sites (e.g., betclic.pt) acknowledge that goals in the second half are more likely than in the first half of football matches.

Ultimately, we can also observe a sharp increase in VAEP scores as the games approach each half, which we also attribute to the presence of riskier strategies. For instance, at the stoppage time of the second half, with nothing more to lose, a goalkeeper from a losing team can move forward in the field for a set piece which is a high-risk decision that increases the odds of a goal being scored (most of the times in its own goal).

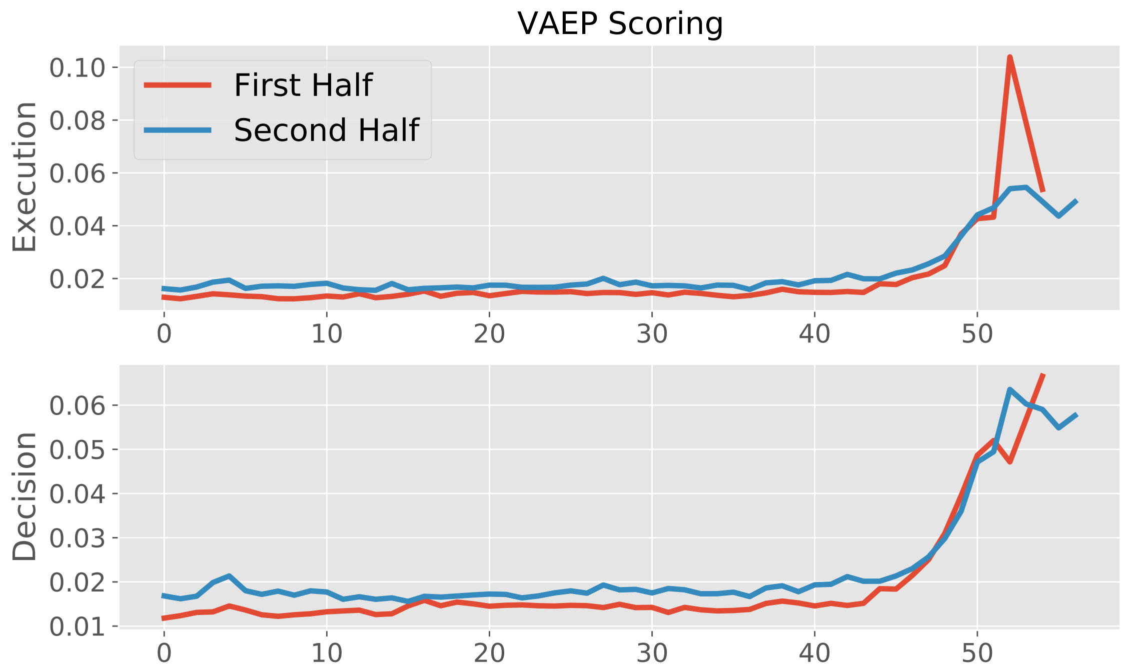

Figure 5.5: VAEP scoring execution and decision values across game time

Regarding space, we can observe from Figure 5.6 that the actions with higher VAEP scores occur near the two penalty areas, which is logical. As teams approach the finishing area, the chances for a given action to precede a goal sharply increase, and, as so, the VAEP score also increases. Indeed, Meireles (2021) conducted interviews with subject matter experts (i.e., former football players that were European Champions), which stated that a primary aim for the defenders was to prevent the attackers from getting into the "imaginary triangle" between the interception of the end with the penalty area lines and the space in front of the penalty area. Curiously, such a triangle can be well observed in the Figure if we take the attacker's perspective; the most "dangerous" actions occur within such a triangle.

5.2 Win-Draw-Loss Probabilities During Matches

As observed from the Figure 5.7, we accurately calculated the win-draw-loss probabilities as a given game unfolded. In the left plot, the probability of the home team ('Fiorentina') winning (blue line), drawing (orange line), and losing the game (green line) changes as the match evolves. The peaks in the green correspond to the goals suffered by the home team ('Fiorentina'), which

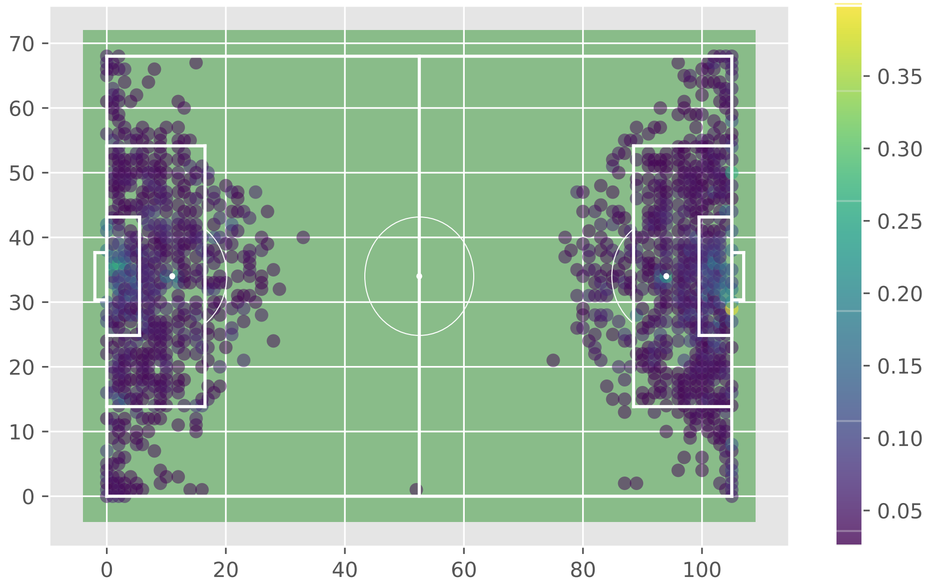

Figure 5.6: VAEP scoring execution variation in space for actions above 0.025

were confirmed by consulting the information about that game ('Fiorentina x Juventus'). In the right plot, we can simply observe that the win-draw-loss model can move from the first to the second half.

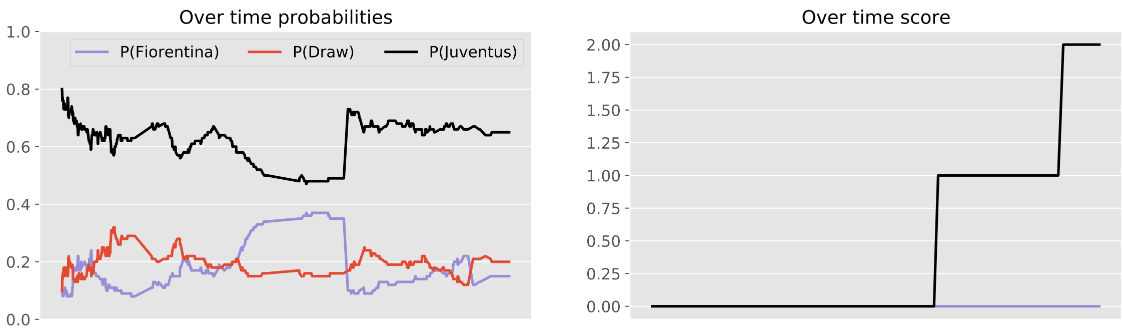

Figure 5.7: For the Fiorentina vs. Juventus 2017/2018 - In the left, win-draw-loss probabilities across time. In the right, moving from the first to the second period of the game

5.3 League Outcome Probabilities

Regarding the estimated outcome probabilities, we can observe from Figures 5.8 and 5.9 that the obtained estimations vary as expected as the season unfold. In the first Figure, we can see, from the perspective of Juventus - the team that eventually won the Championship - how the likelihood of different league outcomes (i.e., becoming a champion, qualifying for the Champions League, qualifying for the Europa League, and being relegated) changed as the season evolved. In the second, we can see how the odds of becoming Champion evolved for the most prominent teams in the Italian league during the 2017-2018 season. Again the behavior of the estimations seems to make sense, particularly when comparing those for Juventus and Napoli - the team that eventually became the second classified. The estimated probability of becoming Champion seems to vary inversely for these two teams.

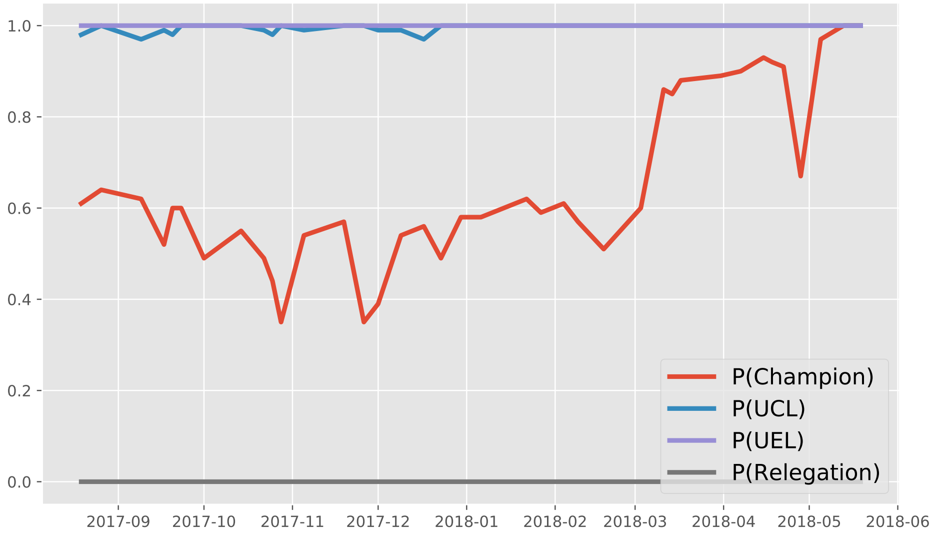

Figure 5.8: Season estimated outcome for Juventus

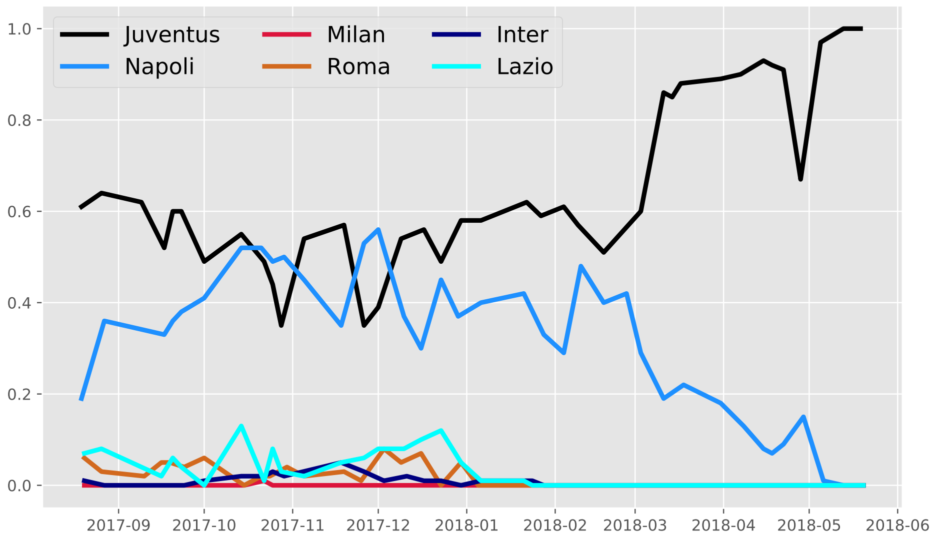

Figure 5.9: Estimated probability of Italian main teams becoming champions

Chapter 6

Modeling

In this section of the work, we report the efforts made to model normative contextual pressure, documenting the several steps that we performed, encompassing data integration, exploratory data analysis, and the theoretical reasoning underlying the modeling effort.

At first, we integrated data from the several data sources that we had. Namely, the calculation of the individual contributions (i.e., the VAEP scores), the win-draw-loss probabilities during matches, and the league outcome probabilities files. Then, we explored the data using bivariate analysis to identify which features made more sense to inform our model of normative contextual pressure model.

Like Bransen et al. (2019) (see Chapter 3), we choose to model normative contextual pressure by considering that it is composed by two distinct types of pressure: the pre-game and the in-game contextual pressure. The pre-game pressure relates to the importance of the game given the statistical outcome of the season for each team, the closeness in strength between each confronting teams, and other contextual factors that may influence the result of a match (e.g., game location: away vs. home). When considered together, these factors may influence the level of pressure a team or player feels before a match begins. Significantly, and unlike Bransen et al. (2019), we considered that the level of confidence of a given player could also be approximated from the average VAEP contribution within a given time frame. According to Lazarus (2000); Lazarus and Folkman (1984); Cruz (1996), the level of confidence, whether the player appraises himself as a more or less competent performer, is an essential determining factor of the perceived pressure.

On the other hand, the in-game mental pressure describes the variation of pressure levels imposed by the game events as a given match unfolds. For instance, a received red card, or a conceded goal in a tight game are events that will likely make the pressure mount for the team receiving the red card or conceding the goal.

6.1 Modeling Pre-Game Normative Contextual Pressure

As referred above, we believe that pre-game normative contextual pressure is influenced by the game's importance for the statistical outcomes of the two teams, the closeness between each team's SPI values, and the players' confidence level. For the closeness factor, following an Elo-based logic, we hypothesized that games between teams with closer SPI values will tend to be perceived by the players as more stressful than games for which that difference was higher since unpredictability about the outcome of the game is also higher. According to Lazarus and Folkman (1984) and Lazarus (2000), unpredictability or uncertainty is a facilitative factor of stress.

Therefore, when modeling pre-game pressure, defined in Equation 6.3, we considered the following arguments:

- **Importance**: the importance of the game for each team attributed to each match, obtained by Equation 4.6.
- **Closeness of Team Strength Values**: how apart are each teams' SPI values, using the normalized difference between the teams' Elo ratings, obtained by Equation 6.1.
- **Confidence Value**: the normalized inverse of the (rolling) average VAEP scores for the past month, obtained by Equation 6.2.

$$Closeness = 1 - norm(|SPI_{home} - SPI_{away}|) \tag{6.1}$$

$$Confidence = 1 - norm(mean_{last\ 30\ days}(VAEP)) \tag{6.2}$$

$$Pre-Game\ Pressure = Importance + Closeness + Confidence \tag{6.3}$$

6.2 Modeling In-Game Normative Contextual Pressure

We believe that in highly disputed or tighter matches, the pressure tends to mount until the end of the match, and the opposite occurs when the winning/losing margin is too big for the losing team to recover and get back into the game. Therefore, we computed a measure of a game's tightness level to model the in-game normative contextual pressure. Our approach took advantage of the win-draw-loss probability estimation and the normalized standard deviation between the three probabilities to check whether the three probabilities were close or apart—the smaller the standard deviation, the tighter the matches. After normalizing the values between 0 and 1, we have:

$$In-Game\ Pressure = 1 - norm(std([P_{win}, P_{draw}, P_{lose}])) \tag{6.4}$$

From this, the Game Normative Contextual Pressure can be obtained from the following equation, and just as in Bransen et al. (2019):

$$NormativeContextualPressure = Pre-Game\ Pressure * In-Game\ Pressure \tag{6.5}$$

The correlation matrix in Figure 6.1 displays the correlation between all the pressure metrics. The fact that the in-game pressure metric is clearly more correlated to the normative contextual pressure metric makes sense to us since the former depends on the win-draw-loss probabilities during the game, which already considers each team's value and the current score. Intuitively, the pre-game pressure metric is a more "theoretical" kind of pressure because it does not acknowledge the unfolding events of a game; likewise, its influence on normative contextual pressure makes sense to be more limited.

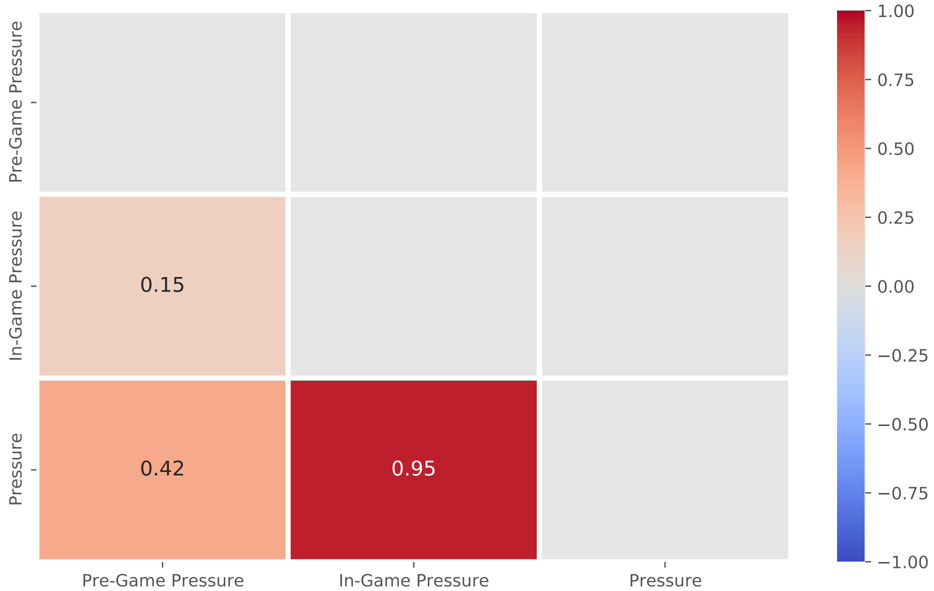

Figure 6.1: Correlation matrix for the several normative contextual pressure metrics

6.3 Discretizing Pressure: defining intervals for low, medium, and high-pressure

Unlike Bransen et al. (2019), which opted for relativist criteria to define the intervals corresponding to different levels of pressure, we opted for a more business-oriented approach. Accordingly, we used more stringent standards instead of simply dividing the distribution so that the 20 percent lower pressure values correspond to the category low-pressure and the 20 percent higher to high-pressure. We believe that high-pressure situations in a soccer match are rarer than 20%, and many actions captured in event streaming data do not involve high-risk situations. For instance, about half of the events recorded in this data set are passes (see Figure 6.2), and many of the passes performed in a game do not imply a high level of risk (e.g., those between two center backs when the team is in an offensive organization). Hence, by inspecting the distributions in a histogram (see Figure 6.3), we discretized pressure by setting the following categories:

- **Low Pressure**: pressure values below percentile 65.

- **Medium Pressure**:pressure values between percentile 65 and 91
- **High Pressure**: pressure values above the percentile 92.

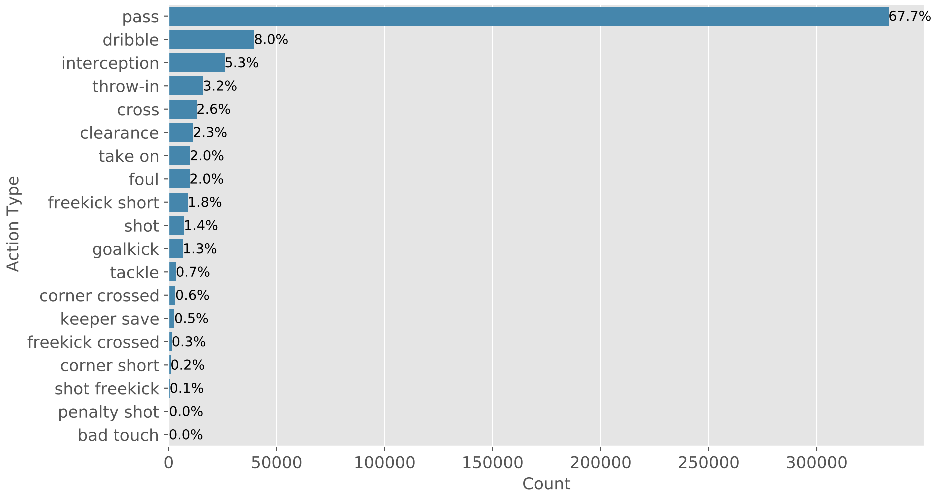

Figure 6.2: Frequency of different actions

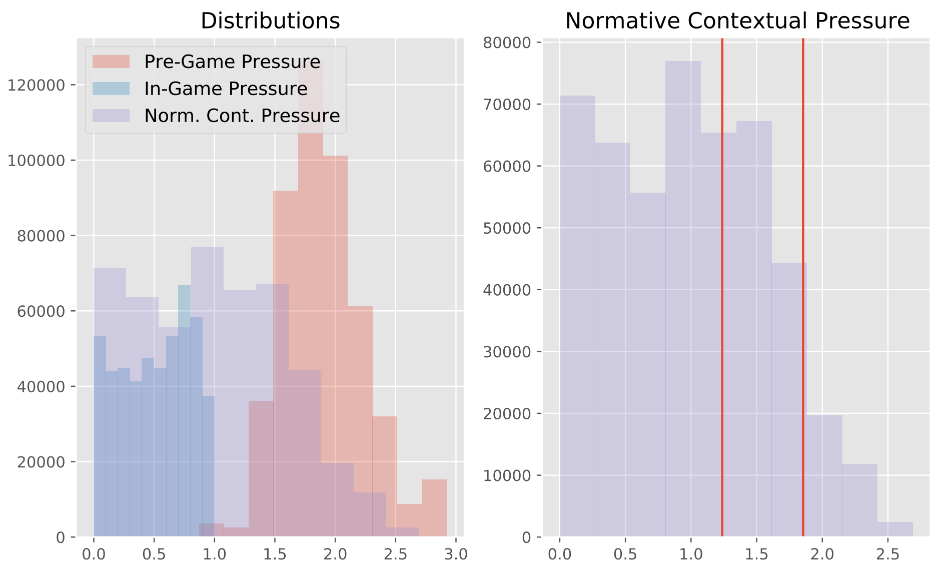

Figure 6.3: In the left, histograms for pre-game, in-game and normative contextual pressure. In the right, normative contextual pressure histograms with vertical lines signalizing the percentile values (65, 92) used for discretization in low, medium and high-pressure situations

As reported in Tables 6.1 and 6.2, the average values for VAEP scoring execution and VAEP decision execution follow the same pattern. They are slightly higher for moderate-pressure situations than low-pressure situations and sharply decrease for high-pressure situations. This pattern narrowly mimicked the Yerkes and Dodson law (Yerkes et al., 1908) for the relationship between stress and performance and was expected. The Yerkes and Dodson law states that the stress and performance relationship follows an inverted U-shaped curve, according to which performance tends to be lower for low and high-stress situations than for moderate (or optimal) stress levels. Furthermore, the results clearly indicate that both execution and decision are impaired in high-pressure situations.

Table 6.1: VAEP scoring execution average values and considered number of instances after defining intervals for low, medium, and high-pressure levels

	VAEP Scoring Execution		
	Low Pressure	Moderate Pressure	High Pressure
Avg Performance	0,00296	0,00299	0,00251
Number of Instances	311035	129199	38282

Table 6.2: VAEP scoring decision average values and considered number of instances after defining intervals for low, medium, and high-pressure levels

	VAEP Scoring Decision		
	Low Pressure	Moderate Pressure	High Pressure
Avg Performance	0,0041	0,00413	0,0038
Number of Instances	311035	129199	38282

Given that event stream data is clearly unbalanced towards attacking actions, we inspected how pressure influenced performance on two types of these actions: take-ons and shots. As can be observed in the Figures 6.4 and 6.5, pressure tends to be detrimental to performance for both actions, especially for the shooting actions. On average, the shooting performance is much lower for high-pressure situations than for lower and moderate-pressure conditions.

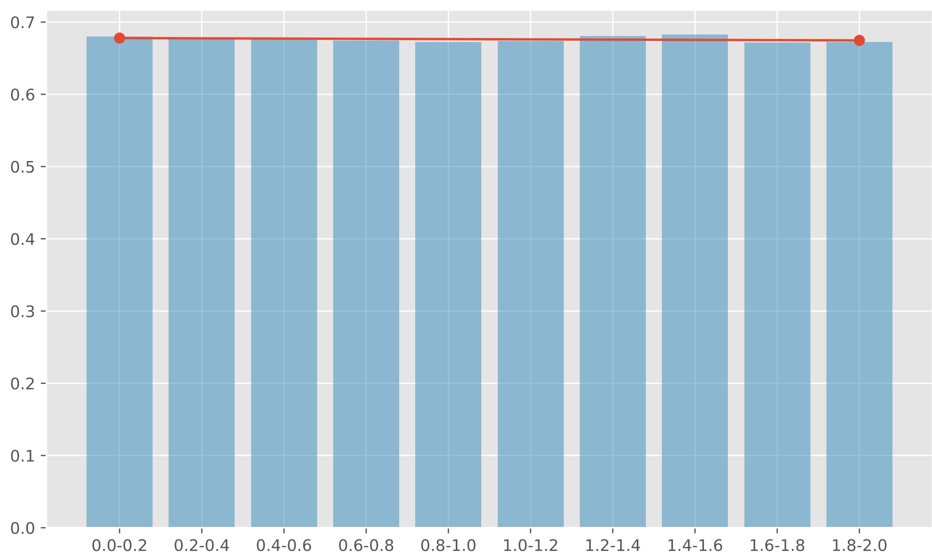

Figure 6.4: VAEP execution scores according to normative contextual pressure for take-ons

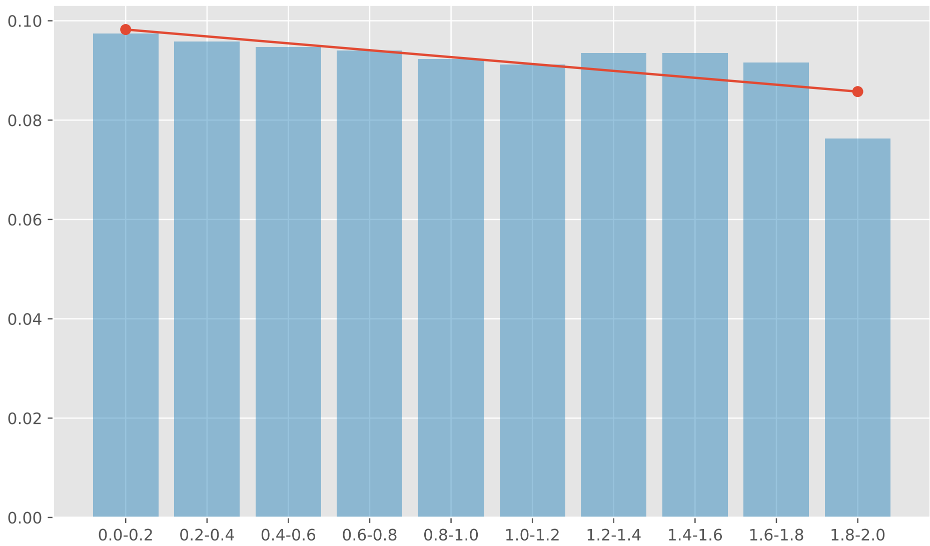

Figure 6.5: VAEP execution scores according to normative contextual pressure for shots

6.4 Exploring the performance under pressure profile of the top scorers and the best players for the target season

To examine the explainability power of our pressure metric, we explored how the top scorers and the best players from Italian Serie A for the 2017-2018 season performed in situations with low, moderate, and high-pressure levels. While always speculative, we can identify different performance profiles by looking at these Figures, particularly when attending to additional information regarding the evolution of each player's career or other relevant developmental and demographic data.

In the case of top scorers (see Figure 6.6), we observe that high pressure seems to be generally detrimental to performance. However, there were two cases in which two players, Higuaín and Quagliarella, performed more valuable actions when the pressure was high. Curiously, these players were two veteran international-level players aged over 30 and 34 years old, respectively. It can be the case (as it is very likely) that players may learn how to cope better with high moments of pressure as their careers unfold. The case of Giovanni Simeone is the opposite; he was a young talented player at 22 years old, and, as pressure mounted, his performances clearly declined. In the last few years, he did not confirm all the potential that he early has shown up. Regarding the best players (see Figure 6.7), we can see that several players have contributed the most to the probability of their team scoring during high-pressure situations. The case in which a decline in high-pressure situations was most evident was that of J. Ilicic. A player whose career in the last years has been affected by psychological issues.

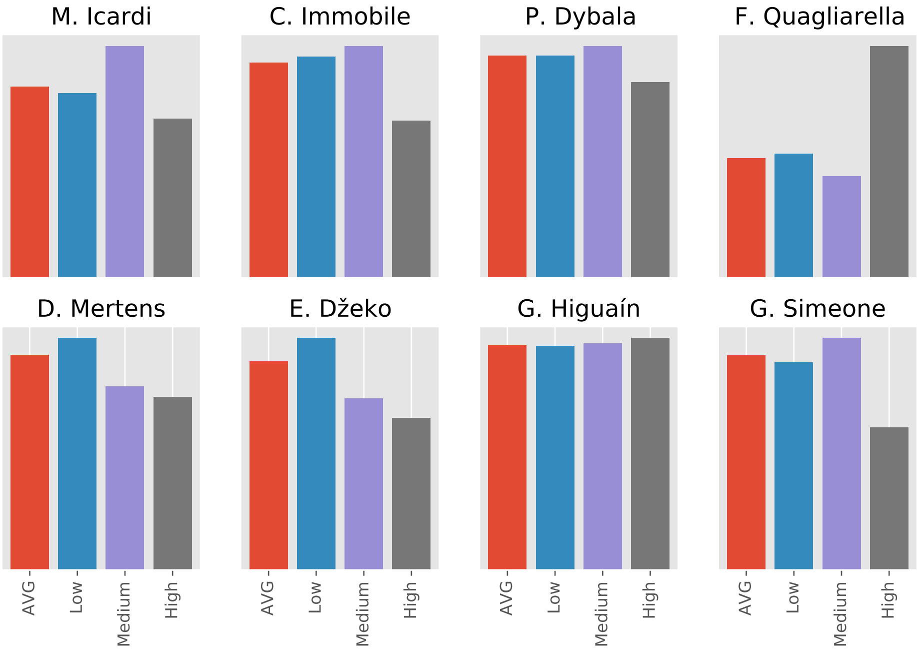

Figure 6.6: VAEP execution scores for the top goal-scorers in the league for low, medium and high-pressure situations

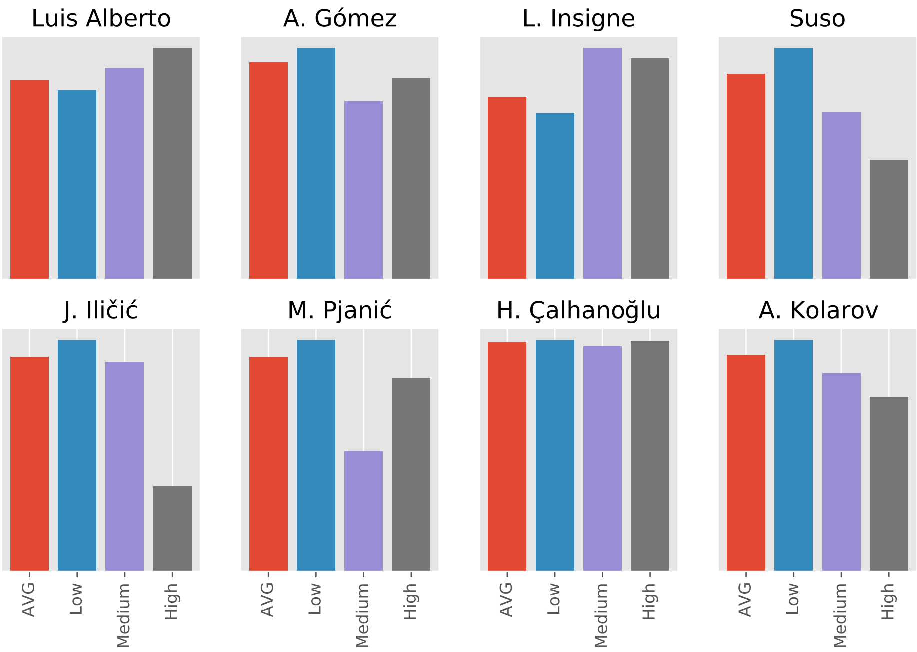

Figure 6.7: VAEP execution scores for the best rated players in the league for low, medium and high-pressure situations

6.5 Use Case: Grouping football players regarding their performance under normative contextual pressure

One of the fundamental processes within football clubs is talent identification and recruitment. However, talent identification is not a trivial task because the concept of talent is itself complex. According to Meylan et al. (2010), rather than the sum of technical, tactical, physical, and psychological skills, a talented football player exhibits the right balance of them. Nevertheless, given all the attention devoted to football, a new player that does not fit the team rapidly becomes noticed and criticized, which inflicts tremendous pressure on all those involved in the talent identification and recruitment process.

Herein, we used our pressure metric to inform player recruitment about the ability of football players to cope with normative contextual stress. Regarding this use case, we have followed an unsupervised machine learning approach to identify groups of distinct players concerning their performance under low, moderate, and high-pressure levels.

Specifically, we have used the K-Means algorithm using the predefined 'sklearn' criteria to cluster football players regarding their performance at low, moderate, and high-pressure levels. We used the elbow method to identify the most appropriate number of clusters (see Figure 6.8), with 5 being the number of clusters indicated by the method. However, we also tested the same algorithm for 3 clusters.

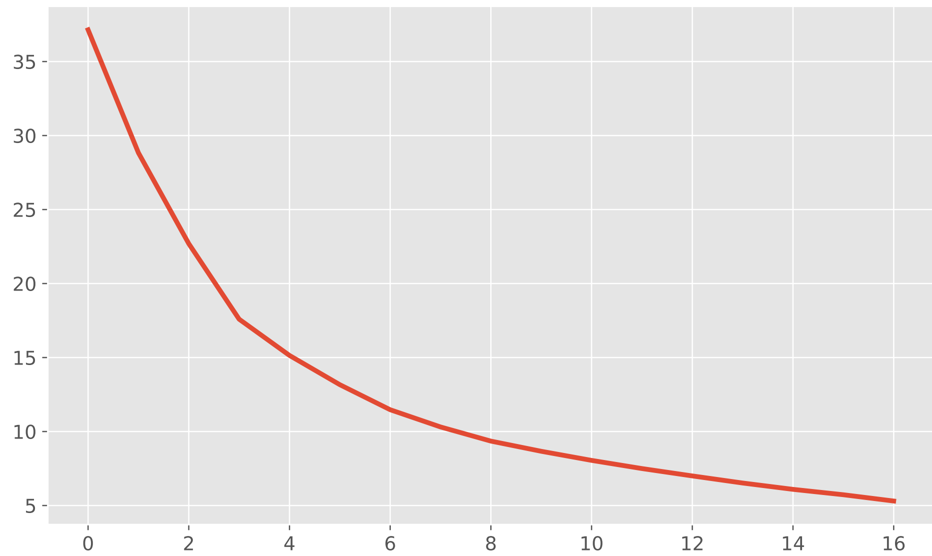

Figure 6.8: The elbow curve for determining the optimal number of clusters

While the Tables 6.3 and 6.4 provide the average values for each cluster given the features: low, moderate, and high pressure, the graphs provide a more direct visualization of the identified profiles. We have more diverse profiles for the 5 clusters scenario, as we can see in Figure 6.9.

Cluster 1 comprises players that perform more valuable actions in low pressure conditions, with similar VAEP scores for both moderate and high-pressure situations. Cluster 2 includes players that perform much better in high-pressure situations and are likely to be substitute players thrown into tight matches that eventually make an assist or score a goal. Cluster 3 contains players with very acceptable scores for both low and moderate levels categories, but that seem to choke under pressure; as pressure, continuously mount their performance level decreases. Cluster 4 contains the players whose performances follow the typical inverted U-shape with better scores on moderate pressure situations. Finally, cluster 5 holds those players that whose performance increases as pressure mounts, with very acceptable values for all the varying levels of pressure.

The three clusters condition has proven to be very limited for identifying the players that cope better with normative contextual pressure as an indicator of (psychological) talent. As seen from the Figure 6.10, each cluster groups the players toward the maximum of each of the categories (i.e., low, moderate, and high-pressure). In Cluster 1 we have players with clearly better performances under high-pressure, in the Cluster 2 the players that perform better under moderate normative stress levels, and, finally, in Cluster 3 the players with higher performance scores for the low pressure condition.

In sum, if we had to recommend players for a club to recruit based on this kind of approach, we would recommend those from Cluster 5 in the five Clusters model. That cluster contains players with a pretty acceptable performance level for low-pressure situations and comparable high values for moderate and high-pressure conditions.

Table 6.3: 5 clusters scenario - frequency and average values for the low, medium and high-pressure

	Cluster 1 n = 59	Cluster 2 n=61	Cluster 3 n=54	Cluster 4 n=46	Cluster 5 n=56
Low	0,9990	0,3860	0,8716	0,3272	0,6650
Moderate	0,3443	0,3187	0,9314	1	0,8421
High	0,4671	1	0,3510	0,3352	0,9134

Table 6.4: 3 clusters scenario - frequency and average values for the low, medium and high-pressure

	Cluster 1 n = 92	Cluster 2 n=99	Cluster 3 n=85
Low	0,4472	0,5643	0,9979
Moderate	0,4875	0,9964	0,5029
High	0,9981	0.4672	0,5168

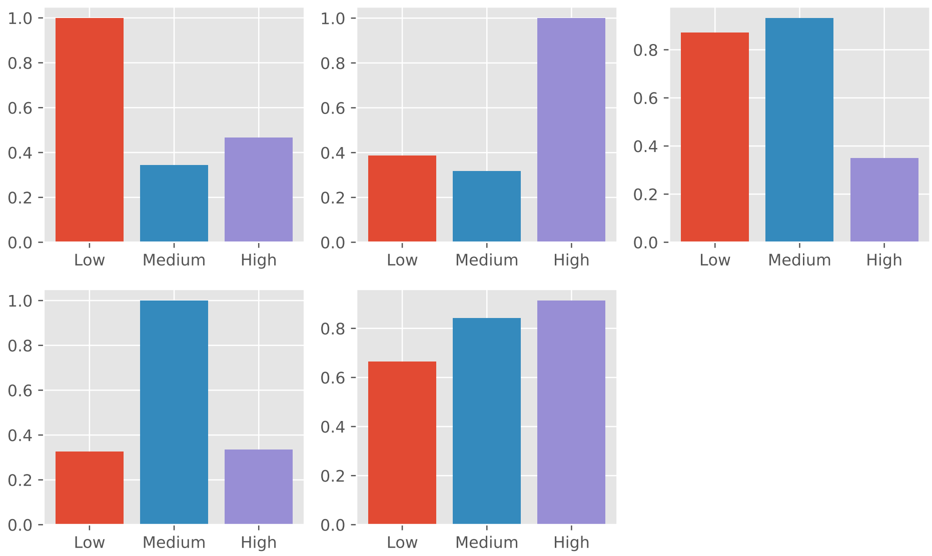

Figure 6.9: 5 clusters scenario - graphical representation of the clusters

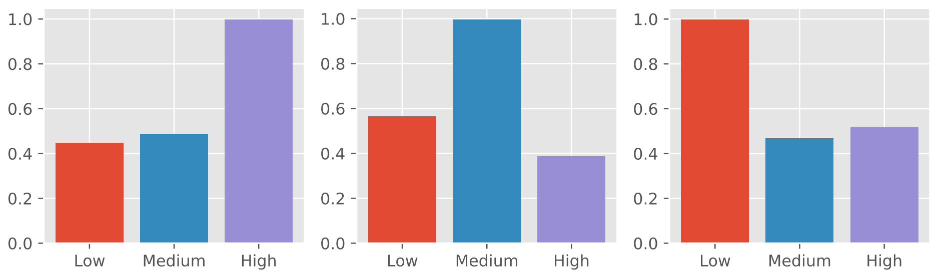

Figure 6.10: 3 clusters scenario - graphical representation of the clusters

Chapter 7

Conclusions

As stated in the Introduction, this thesis pursued three specific aims and one ultimate purpose. Regarding the first aim, we think we provide a fair reflection regarding how Psychology sees football players' performance, particularly regarding the influence of normative contextual pressure and how the psychological processes embedded in stressful generations occur.

By the way, those processes were considered when modeling normative contextual pressure, the second aim. We made a considerable effort of data wrangling and feature engineering to calculate the features required to model normative contextual stress in such a way that it accorded one of the most prominent theories of stress - the Transactional Model of Stress by Lazarus and Folkman (1984). Although future improvements are undoubtedly welcome, our work represents a step ahead of the only known similar work (Bransen et al., 2019).

With the third aim, rather than presenting a solid and polished product, we wanted to demonstrate that we could use raw data (i.e., event streams) to generate valuable knowledge regarding something as abstract as football players' psychological profiles. Furthermore, we wanted to show that such knowledge could inform decision-making at a club level (e.g., talent identification and development). Indeed, our model can be improved, and Machine Learning methods other than unsupervised knowledge can and should be tested in the future.

Regarding the general purpose, we wanted to convince more skeptical people that it is possible to construct knowledge from raw football data through appropriate data mining. To refute Cruyff, we wanted to show that it was possible to detect more abstract things as technique, vision, and even psychological vulnerability to stress from "computer stats" and to use it in a business-informed way. Only time will tell if this work is convincing enough. Notwithstanding, we are already thinking of ways to improve several steps of the method we followed.

7.1 Limitations and Future Work

Despite its merits, this work also has several significant limitations that shall be acknowledged. In the first place, the available data to train the models we used to generate features was limited. We only had data from one season for five different leagues. Thus we needed to use data from competitions other than the target league to use machine learning methods and generate the features. Ideally, we would rather have a historical data set in which we would use data from the past to train the models to be tested in current data. This would prevent several putative cultural biases regarding the natural differences between each league.

Furthermore, it would have enabled us to use different data-mining approaches. For instance, we could have followed a time series paradigm and inspected how vulnerability to stress evolved in more extended time frames. We could even have studied how coping ability changed throughout a player's whole career and confirm/infirm the hypothesis that more experient football players are better than novices at dealing with normative competitive stress.

Due to the extensive time required to treat the raw data and make it manageable to model normative contextual pressure, we lacked the time to properly optimize/tune each model we used during our feature engineering process. In the future, such an optimization/tunning shall be performed using methods such as grid search or Bayesian optimization.

Finally, it would be interesting to investigate how normative contextual pressure relates to subjective measures of stress, such as the threat/challenges perception scales (Cruz (1996). More than providing us with a "ground truth" to evaluate the external validity of our metrics since they may estimate different things, it could be interesting to study in which cases normative pressure correlates with perceived stress and the cases in which that doesn't happen. Do the players performing better in high-pressure situations perceive them as stressful? It is an open question!

Appendix A

Complementary Information

A.1 The Elo Rating System

The ELO rating system was initially developed for chess by Elo Arpad (1978). However, it can be adapted to compute the relative skill level of players or teams in several sports or board games. This system works as a standard scale that measures the cumulative value of any performance of a given competitor. In that sense, and at any given time, it is possible to sort by strength a given pool of competitors.

Elo Arpad (1978) modeled the performance of any given competitor (i.e., individual or team, depending on the situation) as a normally distributed random variable. The same is to say that Elo conceived a player's actual skill as the mean of a random variable for such a player. Considering that, deviations from the mean (up or down) are relatively slow. Looking at the argument Dp from the classic ELO rating formula is easy to understand why.

$$R_p = R_c + D_p \tag{A.1}$$

where:
R_p is the performance rating
R_c is the aggregated (average) competition rating
D_p is the difference based on the percentage score P, which is obtained by the normal curve.

Elo (1978) successfully pictured the operationalization of argument D_p using the figure below. When used on continuous basis (i.e.m updated from observation to observation), the formula to calculate the rating is the following:

$$R_n = R_o + K(W - We) \tag{A.2}$$

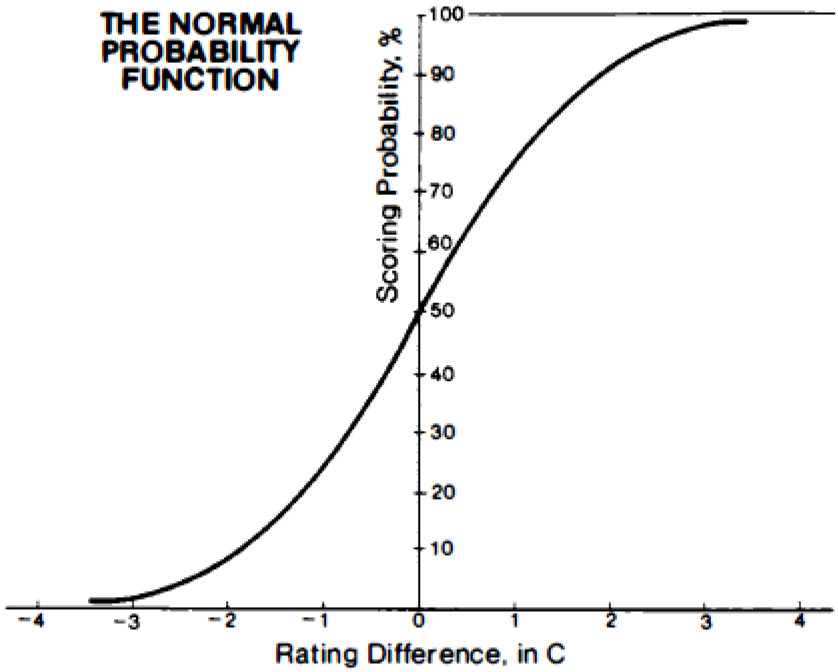

Figure A.1: The Gauss error curve, or standard sigmoid, used to model the relationship between the winning/losing probability and rating differences (Adapted from Elo Arpad (1978))

where:
R_n is the new rating after an observation (e.g., a game)
R_0 is the pre-observation rating (e.g., before the game)
K is the rating point value of a favourable observation (i.e., a scoring or a winning)
W is the result of the observation. Each win counting 1 and each draw $1/2$
We is the expected result based on R_o

The ELO rating system was already adapted to make football predictions -the same is to say: to predict the probability of a team winning, drawing, or losing a game (Hvattum and Arntzen, 2010). The authors explain a variation of the ELO rating for ball games (e.g., football), which allows the coefficient K to be updated by goal difference instead of win-draw-loss information, thus rewarding a 5-1 win more strongly than a 1-0 win. Formally, the equation 2.6, suffers the following transformation:

$$R_n = R_0 + K(1+\delta)^{\lambda} \tag{A.3}$$

where:
R_n is the new rating after an observation (e.g., a game)
R_0 is the pre-observation rating (e.g., before the game)
K is the rating point value of a favourable observation (i.e., a scoring or a winning)
δ is the absolute goal difference
K and $\lambda > 0$

Appendix B

Public WyScout data set

	event_id	game_id	period_id	milliseconds	team_id	player_id
696	253669083	2576335	1	2507022.486	3161	116349
231	253668585	2576335	1	711239.457	3162	265865
729	253669096	2576335	1	2571824.815	3162	346908
61	253668375	2576335	1	172233.939	3161	3431
1071	253670389	2576335	2	775271.557	3162	0
167	253668520	2576335	1	516821.288	3162	40806
713	253669103	2576335	1	2545612.112	3161	138408

	event_id	type_id	type_name	subtype_id	subtype_name
696	253669083	8	Pass	85	Simple pass
231	253668585	1	Duel	12	Ground defending duel
729	253669096	1	Duel	12	Ground defending duel
61	253668375	8	Pass	85	Simple pass
1071	253670389	1	Duel	11	Ground attacking duel
167	253668520	8	Pass	80	Cross
713	253669103	8	Pass	85	Simple pass

	event_id	positions	tags
696	253669083	[{’y’: 56, ’x’: 44}, {’y’: 26, ’x’: 35}]	[{’id’: 1801}]
231	253668585	[{’y’: 37, ’x’: 67}, {’y’: 39, ’x’: 68}]	[{’id’: 702}, {’id’: 1801}]
729	253669096	[{’y’: 88, ’x’: 26}, {’y’: 93, ’x’: 33}]	[{’id’: 703}, {’id’: 1801}]
61	253668375	[{’y’: 34, ’x’: 30}, {’y’: 20, ’x’: 48}]	[{’id’: 1801}]
1071	253670389	[{’y’: 9, ’x’: 88}, {’y’: 33, ’x’: 69}]	[{’id’: 701}, {’id’: 1802}]
167	253668520	[{’y’: 3, ’x’: 77}, {’y’: 67, ’x’: 91}]	[{’id’: 402}, {’id’: 801}]
713	253669103	[{’y’: 75, ’x’: 45}, {’y’: 56, ’x’: 49}]	[{’id’: 1801}]

Table B.1: Sample of untransformed Public Wyscout data set

Appendix C

Transformed WyScout data set

	game_id	period_id	time_seconds	team_id	player_id
0	2576335	1	2507.022486	3161	116349
1	2576335	1	1339.628213	3161	3431
2	2576335	1	172.233939	3161	3431
3	2576335	1	516.821288	3162	40806
4	2576335	1	2545.612112	3161	138408

	start_x	start_y	end_x	end_y
0	58.80	38.08	68.25	17.68
1	68.25	17.68	73.50	23.12
2	73.50	23.12	54.60	13.60
3	80.85	65.96	95.55	22.44
4	57.75	51.00	53.55	38.08

	game_id	original_event_id	bodypart_id	type_id	result_id	action_id
0	2576335	253669083	0	0	1	0
1	2576335	NaN	0	21	1	1
2	2576335	253668375	0	0	1	2
3	2576335	253668520	0	1	1	3
4	2576335	253669103	0	0	1	4

Table C.1: Sample of transformed Public Wyscout data set

References

Gennady Andrienko, Natalia Andrienko, Guido Budziak, Jason Dykes, Georg Fuchs, Tatiana von Landesberger, and Hendrik Weber. Visual analysis of pressure in football. *Data Mining and Knowledge Discovery*, 31(6):1793–1839, 2017.

Gabriel Anzer and Pascal Bauer. A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). *Frontiers in Sports and Active Living*, 3:624475, 2021.

Adrià Arbués-Sangüesa, Adrián Martín, Javier Fernández, Gloria Haro, and Coloma Ballester. Towards soccer pass feasibility maps: the role of players' orientation. *Journal of Sports Sciences*, 39(24):2772–2785, 2021.

Pascal Bauer and Gabriel Anzer. Data-driven detection of counterpressing in professional football. *Data Mining and Knowledge Discovery*, 35(5):2009–2049, 2021.

Roy F Baumeister. Choking under pressure: self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of personality and social psychology*, 46(3):610, 1984.

Alina Bialkowski, Patrick Lucey, Peter Carr, Iain Matthews, Sridha Sridharan, and Clinton Fookes. Discovering team structures in soccer from spatiotemporal data. *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2596–2605, 2016.

Lotte Bransen, Pieter Robberechts, Jan Van Haaren, and Jesse Davis. Choke or shine? quantifying soccer players' abilities to perform under mental pressure. In *Proceedings of the 13th MIT Sloan Sports Analytics Conference*, pages 1–25. MIT SLOAN, 2019.

Ulf Brefeld, Jan Lasek, and Sebastian Mair. Probabilistic movement models and zones of control. *Machine Learning*, 108(1):127–147, 2019.

Joel Brooks, Matthew Kerr, and John Guttag. Developing a data-driven player ranking in soccer using predictive model weights. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 49–55, 2016a.

Joel Brooks, Matthew Kerr, and John Guttag. Using machine learning to draw inferences from pass location data in soccer. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(5):338–349, 2016b.

Charles S Carver and Michael F Scheier. *Attention and self-regulation: A control-theory approach to human behavior*. Springer Science & Business Media, 2012.

Hyeonah Cho, Hyunyoung Ryu, and Minseok Song. Pass2vec: Analyzing soccer players' passing style using deep learning. *International Journal of Sports Science & Coaching*, 17(2):355–365, 2022.

Noam Chomsky. *Aspects of the Theory of Syntax*, volume 11. MIT press, 2014.

José Fernando A Cruz. *Stress e ansiedade na competição desportiva: natureza, efeitos e avaliação*. SHO-Sistemas Humanos e Organizacionais, Lda, 1996.

Tom Decroos and Jesse Davis. Player vectors: Characterizing soccer players' playing style from match event streams. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 569–584. Springer, 2019.

Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1851–1861, 2019.

Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Vaep: an objective approach to valuing on-the-ball actions in soccer. In *Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI-20*, pages 4696–4700. International Joint Conferences on Artificial Intelligence Organization, 2020a.

Tom Decroos, Jesse Davis, et al. Valuing on-the-ball actions in soccer: a critical comparison of xt and vaep. In *Proceedings of the AAAI-20 Workshop on Artifical Intelligence in Team Sports*, pages 1–8. AI in Team Sports Organising Committee, 2020b.

Tom Decroos, Maaike Van Roy, and Jesse Davis. Soccermix: representing soccer actions with mixture models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 459–474. Springer, 2020c.

Jordi Duch, Joshua S Waitzman, and Luís A Nunes Amaral. Quantifying the performance of individual players in a team activity. *PloS one*, 5(6):e10937, 2010.

E Elo Arpad. The rating of chess players. *Past and Present. BT Batsford, Ltd*, 1978.

Mica R Endsley. Toward a theory of situation awareness in dynamic systems. In *Situational awareness*, pages 9–42. Routledge, 2017.

K Anders Ericsson, Ralf T Krampe, and Clemens Tesch-Römer. The role of deliberate practice in the acquisition of expert performance. *Psychological review*, 100(3):363, 1993.

K Anders Ericsson, Robert R Hoffman, Aaron Kozbelt, and A Mark Williams. *The Cambridge handbook of expertise and expert performance*. Cambridge University Press, 2018.

Javier Fernandez and Luke Bornn. Wide open spaces: A statistical technique for measuring space creation in professional soccer. In *Sloan sports analytics conference*, volume 2018, 2018.

Javier Fernández, Luke Bornn, and Daniel Cervone. A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions. *Machine Learning*, 110(6):1389–1427, 2021.

Floris R Goes, Matthias Kempe, Laurentius A Meerhoff, and Koen APM Lemmink. Not every pass can be an assist: a data-driven model to measure pass effectiveness in professional soccer matches. *Big data*, 7(1):57–70, 2019.

Daniel F Gucciardi and James A Dimmock. Choking under pressure in sensorimotor skills: Conscious processing or depleted attentional resources? *Psychology of Sport and Exercise*, 9(1): 45–59, 2008.

Laszlo Gyarmati and Rade Stanojevic. Qpass: a merit-based evaluation of soccer passes. *arXiv preprint arXiv:1608.03532*, 2016.

Jan De Houwer. Why the cognitive approach in psychology would profit from a functional approach and vice versa. *Perspectives on Psychological Science*, 6(2):202–209, 2011.

Lars Magnus Hvattum and Halvard Arntzen. Using elo ratings for match result prediction in association football. *International Journal of forecasting*, 26(3):460–470, 2010.

Frederick James. Monte carlo theory and practice. *Reports on progress in Physics*, 43(9):1145, 1980.

Matthias Kempe, Floris R Goes, and Koen APM Lemmink. Smart data scouting in professional soccer: Evaluating passing performance based on position tracking data. In *2018 IEEE 14th International Conference on e-Science (e-Science)*, pages 409–410. IEEE, 2018.

Dirk P Kroese, Tim Brereton, Thomas Taimre, and Zdravko I Botev. Why the monte carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6):386–392, 2014.

Richard S Lazarus. How emotions influence performance in competitive sports. *The sport psychologist*, 2000.

Richard S Lazarus and Susan Folkman. *Stress, appraisal, and coping*. Springer publishing company, 1984.

Vittorio Lingiardi and Nancy McWilliams. *Psychodynamic diagnostic manual: PDM-2*. Guilford Publications, 2017.

Daniel Link, Steffen Lang, and Philipp Seidenschwarz. Real time quantification of dangerousity in football using spatiotemporal tracking data. *PloS one*, 11(12):e0168768, 2016.

Guiliang Liu, Yudong Luo, Oliver Schulte, and Tarak Kharrat. Deep soccer analytics: learning an action-value function for evaluating soccer players. *Data Mining and Knowledge Discovery*, 34 (5):1531–1559, 2020.

Patrick Lucey, Alina Bialkowski, Mathew Monfort, Peter Carr, and Iain Matthews. quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data, 2015.

Michael J Mahoney and Michael H Craine. The changing beliefs of psychotherapy experts. *Journal of Psychotherapy Integration*, 1(3):207, 1991.

David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.

Luís Jorge Machado da Cunha Meireles. *O papel da "consciência da situação" na tomada de decisão em situação de jogo por parte de jogadores de futebol da elite mundial*. PhD thesis, Universidade do Minho, 2021.

Tiago Mendes-Neves, Luís Meireles, and João Mendes-Moreira. Valuing players over time, 2022.

Cesar Meylan, John Cronin, Jon Oliver, and Michael Hughes. Talent identification in soccer: The role of maturity status on physical, physiological and technical characteristics. *International Journal of Sports Science & Coaching*, 5(4):571–592, 2010.

Walter Mischel and Yuichi Shoda. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, 102(2):246, 1995.

João Moreira, Andre Carvalho, and Tomás Horvath. *A general introduction to data analytics*. John Wiley & Sons, 2018.

Luca Pappalardo, Paolo Cintia, Paolo Ferragina, Emanuele Massucco, Dino Pedreschi, and Fosca Giannotti. Playerank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(5):1–27, 2019.

JOHN S Raglin. Anxiety and sport performance. *Exercise and sport sciences reviews*, 20:243–243, 1992.

Robert Rein and Daniel Memmert. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*, 5(1):1–13, 2016.

Robert Rein, Dominik Raabe, and Daniel Memmert. "which pass is better?" novel approaches to assess passing effectiveness in elite soccer. *Human movement science*, 55:172–181, 2017.

Colin Shearer. The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22, 2000.

Nathan Smith, Damien Tessier, Yannis Tzioumakis, Eleanor Quested, Paul Appleton, Philippe Sarrazin, Athanasios Papaioannou, and Joan L Duda. Development and validation of the multidimensional motivational climate observation system. *Journal of sport and exercise psychology*, 37(1):4–22, 2015.

Ronald E Smith, Frank L Smoll, and Earl Hunt. A system for the behavioral assessment of athletic coaches. *Research Quarterly. American Alliance for Health, Physical Education and Recreation*, 48(2):401–407, 1977.

William Spearman. Beyond expected goals. In *Proceedings of the 12th MIT sloan sports analytics conference*, pages 1–17, 2018.

Silvan Steiner, Stephan Rauh, Martin Rumo, Karin Sonderegger, and Roland Seiler. Outplaying opponents—a differential perspective on passes using position data. *German journal of exercise and sport research*, 49(2):140–149, 2019.

Łukasz Szczepański and Ian McHale. Beyond completion rate: evaluating the passing ability of footballers. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2): 513–533, 2016.

Karl Tuyls, Shayegan Omidshafiei, Paul Muller, Zhe Wang, Jerome Connor, Daniel Hennes, Ian Graham, William Spearman, Tim Waskett, Dafydd Steel, et al. Game plan: What ai can do for football, and what football can do for ai. *Journal of Artificial Intelligence Research*, 71:41–88, 2021.

A Mark Williams and Tom Reilly. Talent identification and development in soccer. *Journal of sports sciences*, 18(9):657–667, 2000.

Rüdiger Wirth and Jochen Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39. Manchester, 2000.

Robert Mearns Yerkes, John D Dodson, et al. The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, 1908.