



UPPSALA
UNIVERSITET

UPTEC F 22032

Examensarbete 30 hp

Juni 2022

Avatar Playing Style

From analysis of football data to recognizable playing styles

Jakob Edberger Persson
Emil Danielsson



Abstract

Football analytics is a rapid growing area which utilizes conventional data analysis and computational methods on gathered data from football matches. The results emerging out of this can give insights of performance levels when it comes to individual football players, different teams and clubs. A difficulty football analytics struggles with daily is to translate the analysis results into actual football qualities and knowledge which the wider public can understand. In this master thesis we therefore take on the ball event data collected from football matches and develop a model which classifies individual football player's playing styles, where the playing styles are well known among football followers. This is carried out by first detecting the playing positions: 'Strikers', 'Central midfielders', 'Outer wingers', 'Full backs', 'Centre backs' and 'Goalkeepers' using K-Means clustering, with an accuracy of 0.89 (for Premier league 2021/2022) and 0.84 (for Allsvenskan 2021). Secondly, we create a simplified binary model which only classifies the player's playing style as "Offensive"/"Defensive". From the bad results of this model we show that there exist more than just these two playing styles. Finally, we use an unsupervised modelling approach where Principal component analysis (PCA) is applied in an iterative manner. For the playing position 'Striker' we find the playing styles: 'The Target', 'The Artist', 'The Poacher' and 'The Worker' which, when comparing with a created validation data set, give a total accuracy of 0.79 (best of all positions and the only one covered in detail in the report due to delimitations).

The playing styles can, for each player, be presented visually where it is seen how well a particular player fits into the different playing styles. Ultimately, the results in the master thesis indicates that it is easier to find playing styles which have clear and obvious on-the-ball-actions that distinguish them from other players within their respective position. Such playing styles, easier to find, are for example "The Poacher" and "The Target", while harder to find playing styles are for example "The Box-to-box" and "The Inverted". Finally, conclusions are that the results will come to good use and the goals of the thesis are met, although there still exist a lot of improvements and future work which can be made.

Developed models can be found in a simplified form on the GitHub repository: <https://github.com/Sommarro-Devs/avatar-playing-style>. The report can be read stand-alone, but parts of it are highly connected to the models and code in the GitHub repository.

Teknisk-naturvetenskapliga fakulteten

Uppsala universitet, Utgivningsort Uppsala

Handledare: Ola Lidmark Eriksson Ämnesgranskare: David J.T.Sumpter

Examinator: Tomas Nyberg

Från analys av fotbollsdata till igenkännbara spelstilar

I skuggan av en alltmer digitaliserad värld har analys av fotbollsdata kommit att bli ett av de viktigaste verktygen i fotbollsklubbars, och andra fotbolls-intressenters, jakt på konkurrensfördelar och framgång. Inom analys av fotbollsdata, eller “Football Analytics” som är den internationella termen som används, ges utrymme för detaljerad statistik och fördjupningar som baseras på matematik och maskininläring. En av de största utmaningarna för “Football Analytics” är översättningen av alla statistiska mått, och deras underliggande ekvationer och algoritmer, till fotbollskvalitéer som är igenkännbara och relaterar till vad som sker ute på fotbollsplanen. En del av den här problematiken hanteras med en nyutvecklad modell som omvandlar fotbollsdata till igenkännbara spelstilar hos olika fotbollsspelare.

Det svenska företaget Football Analytics Sweden tillhandahåller via PlayMaker.AI Sveriges ledande plattform för fotbollsanalys. Plattformen innehåller omfattande verktyg kring scouting, matchanalys och spelarutveckling. En stor del av plattformen utgörs av en spelardatabas innehållandes samtliga fotbollsspelare i de europeiska toppligorna, samt ända ned till Division 1 i Sverige.

Inom en snar framtid kommer den uppdaterade PlayMaker.AI plattformens verktyg kunna kompletteras med tydliga grafiska avatarer och ikoner som representerar de olika spelstilarna en spelare kan ha. Spelstilarna har beräknats fram med en analysmodell som objektivt urskiljer varierande mönster bland olika spelare i den underliggande fotbollsdatan. Jämförelse mellan påvisade spelstilar från modellen mot vedertagna referensspelstilar har gett lovande resultat, i form av god överensstämmelse. I skrivandes stund mynnar modellen ut i 16 olika spelstilar som har gått att identifiera utifrån vissa satta kriterier.

Spelare som indikerar en särskilt stor överensstämmelse är central-offensiva spelare, så kallade ‘Strikers’. En av de framtagna spelstilarna som kanske är den mest igenkännbara är “The Artist”, eller på svenska “Artisten”. “Artisten” utstrålar en elegans som är svår att missa, med en teknik olik någon annan spelare i laget. Kända spelare som kan associeras med “The Artist” är till exempel: Messi, Maradona och Marta.



Att dessa ovan nämnda spelare är tydliga artister behöver man ingen avancerad maskininlärningsmodell för att avgöra; det vet gemene fotbollsintresserad person. Men, vilka än så länge oupptäckta artister finns det i våra lägre divisioner, eller i ligor som kanske inte får en så stor medial uppmärksamhet? Det här, och mycket annat, kan PlayMaker.AI hjälpa till med att besvara på ett helt automatiserat och datadrivet tillvägagångssätt med den nya analysmodellen.

Distribution of work between authors

Overall the work was distributed in a good way with both of us contributing the same amount, and with great collaboration. In early stages of the model development tight teamwork was carried out to come up with different approaches to solving the problem at hand. When the idea stage of the project was finished, some of the work was divided between us, the authors. Jakob has been responsible for programming the pilot model to Model v.2 and the position detection model with visualisations tools for the data analyzes (scree plots, weight patterns, etc.). Emil has been responsible for the development of Model v.1, the final version of Model v.2 with resulting Spider visualisations and the creations of the Jupyter Notebooks. While this was the responsibility distribution, collaboration with continued talks about different parts of the code were held and carried out throughout the programming process/model development.

When it comes to the writing of the report we also worked closely with each other. Here, Jakob wrote most of the Abstract, Introduction, Data background, Modelling approach and Conclusions. Emil wrote most of the Theory, Model explanations/Methods and Discussion. However, as the writing of the report was done in close collaboration, and with an always-open dialog, both of us contributed to every Section to some extent, and are thus both responsible for the report as a whole.

Glossary

- ◇ **Danger zone** - The area of a football pitch from which most danger is created (in terms of shoots and passes leading to goals).
- ◇ **Key pass** - A pass which creates a "good" goal scoring opportunity for a teammate.
- ◇ **Progressive carry** - A continuous ball carry/control from a football player which takes the ball significantly closer to the opponent goal.
- ◇ **Long ball** - A ball kicked/played longer than 30 meters.
- ◇ **The box** - The rectangular area of a football pitch in front of a goal. Also referred to as the penalty area.
- ◇ **Touchline** - The longer sideline of a football pitch.
- ◇ **Goal line** - The shorter sideline of a football pitch, where the goal is placed.
- ◇ **Spider** - A type of radar plot which is common in football analytics to visualize and highlight statistics.
- ◇ **Unsupervised model** - A machine learning model which analyzes and from that cluster/associate unlabeled data. In other words such models try to find patterns and correlation in the data unsupervised (without human intervention).
- ◇ **Supervised model** - A machine learning model which uses sets of labeled (known output from given input) data, often referred to as training data, that defines the models. The results of the model are supervised in the sense of them being output in the same form as the labeled training data.
- ◇ **KPI (Key Performance Indicator)** - Performance measurement of a certain type of action, or ability, connected to what is happening on the football pitch.

Contents

1	Introduction	6
1.1	Background	6
1.2	Aim and Goals	6
1.3	Definitions and Assumptions	7
1.4	Delimitations	7
1.5	Hypothesis	8
1.6	Disposition	10
2	Theory	11
2.1	Normalization techniques	11
2.2	Principal component analysis	11
2.3	Possession adjustment	12
2.4	Share KPIs	12
2.5	Quantile classification	12
2.6	K-Means clustering	12
2.7	Evaluation metrics	13
3	Method and Model development	15
3.1	Data background	15
3.2	Modelling approach	16
3.3	Position detection model	17
3.4	Model v.1	19
3.5	Model v.2	21
4	Results	23
4.1	Position detection model	23
4.2	Model v.1	24
4.3	Model v.2	24
5	Discussion	30
5.1	Position detection model	30
5.2	Model v.1	30
5.3	Model v.2	30
5.4	Future work	32
6	Conclusions	34
	Appendix A	37
1	KPI_org	37

2	KPI_new	38
Appendix B		40
1	Offensive actions	40
2	Defensive actions	40
Appendix C		41
1	Results - Position detection model	41
1.1	Existing positional data/models	41
1.2	New Model	43
2	Results - Model v.1	44
3	Results - Model v.2	45

1 Introduction

1.1 Background

Football is more than just the biggest sport in the world. The interest surrounding events such as the World Cup, the European Championship, the Champions League and other big-league matches are incredible when it comes to fan interaction and coverage by the media. Out of all this, a big industry has grown which, so far, has shown no sign of waning or slowing down in growth [1]. For professional football clubs to succeed in this increasingly competitive environment each club needs to make sure that they smart and effectively allocate their resources in order to maximise their performance. With this the use of data for performance analysis of individual football players and teams has become a big area of focus amongst professional football clubs, as well as the covering media and other parties involved in professional football [2].

The company Football Analytics Sweden provides via PlayMaker.AI Sweden's leading platform for football analysis [3]. The platform contains extensive tools for scouting, match analysis and player development. Users of the platform include professional clubs, scouts and football associations. A major challenge for the platform, and the area of football analytics as a whole, is to translate and transform the actual data analysis to football qualities that all users understand and can relate to in regards to what is taking place on the pitch. As part of the development work around this challenge this project and master thesis takes its starting point, with the objective to characterize the player's (available at PlayMaker.AI) playing styles and presenting it for the platform users.

The benefits with having information and knowing about a player's playing style are several. From professional clubs point of view it gives advantages when it comes to the following:

Scouting possible transfers - In recruitment of new players it is important to know if, and how well, a certain prospect fits into a certain playing style profile that a club is looking to recruit for.

Scouting opponents - In match preparations it can help knowing what type of players the opponents has in order to set up tactics and prepare your own players for what they will face.

Monitoring players and development - When managing your own players it can be of great use to see if you are playing them in their "correct" position and/or role. This can also help to identify which players that can step in and replace other players that are out because of injury or suspension for example.

For Football Analytics Sweden to be able to present such player information on its platform PlayMaker.AI would be of great value and increase the platforms usability even more, would help to ensure the leading role when it comes to football analytics.

1.2 Aim and Goals

The aim with this master thesis is to develop a model that classifies an individual football player's playing style based solely on his or her event data collected from football matches he or she has participated in. In other words, we attempt to classify and characterise a footballer's playing style in a fully data-driven and objective manner.

Specific goals which this master thesis intend to achieve are the following:

- Developed and created model can be used on any football player available on PlayMaker.AI.
- From results the football player's computed playing styles should be presented visually. For example in form of, or combination of:
 - Graphical Avatar.
 - Tags.
 - Spider.
- It should be possible to search for players with a particular playing style.
- The resulting visualisation should show how well a particular player fits into different playing styles and/or positions.

1.3 Definitions and Assumptions

Since a playing style is a highly subjective concept we need a definition of what we in this master thesis consider a player's playing style to be. The definition we use is as follows:

Definition 1 (Playing style). A football player's playing style is defined to be how he or she acts in situations with and around the ball in a football match. The playing style is also characterised by in which areas of the pitch his or her ball actions tend to occur. Following this reasoning, the playing style of multiple players can be distinguished by the different actions they take in similar situations and the outcomes of those actions.

The definition is based on the following assumptions we make:

Assumption 1 (Playing style). A player's playing style does not change within the time period under investigation in the present study.

Assumption 2 (Playing style). Individual differences, but not gender differences, may constitute a basis for differing playing styles.

Assumption 3 (Playing style). A player's playing style depends only on actions with and around the ball.

Assumption 4 (Playing style). Although each player has only one playing style, how well he or she fits into other playing styles may be measured and discussed.

Assumption 5 (Playing style). A player's playing style can be determined/labelled (see Section 3.1.5 for an explanation) subjectively from observing him or her playing in football matches.

Assumption 6 (Playing style). A player's playing style directly follows from the playing position (see Definition 2) he or she plays in. In other words, a "defender" can only have a playing style existing for that defending position (here the position "defender" is made up as an example).

Furthermore as stated in Definition 1 a playing style is partly dependent on the location of a player's actions. In other words, it is position dependent. Hence, in order to identify different playing styles, we first need to consider and identify different playing positions. This follows naturally from how it for example would not make sense to compare playing styles of a defender with a striker, as the main objective for a striker is to score goals and for the defender to prevent the opponent from scoring goals. A player's playing position is defined to be as follows:

Definition 2 (Playing position). A football player's playing position is where on the football pitch the player has been assigned to play in. In other words, a playing position can be regarded as a place on the pitch where the player is mostly located in for strategical purposes.

Assumptions which are made following this are the following:

Assumption 7 (Playing position). There exist six different playing positions.

Assumption 8 (Playing position). Each player only plays in, and belongs to, one playing position.

Assumption 9 (Playing position). Left or right does not matter in terms of playing position, this follows from how the football pitch is symmetric.

1.4 Delimitations

Delimitations that applies to the master thesis are the following:

- We do not take into consideration how a player acts in situations without the ball (we do not have access to this type of data, see Section 3.1).
- We do not take a player's physique or mentality into consideration.
- Goalkeepers are disregarded/excluded, but still need to be identified in order for them to be removed and not interfere.
- Allsvenskan 2021 and Premier League 2021/2022 (up until 2022-02-23) are the only seasons considered in terms of data. Hence they build up the time period this master thesis considers. But the models developed should still work for different seasons and time periods as well.
- No external validation data is considered.

1.5 Hypothesis

Somewhat unusual for a report like this (covering football analytics from a computational science point of view) is to include a hypothesis. Our belief is that the hypothesis in a good way represents both what we want to achieve, and by that concretizes the aims and goals, as well as what we believe is possible to achieve, given the available data and resources (presented in Section 3.1). The hypothesis acts as a common thread throughout the report and will be continuously evaluated during the project work. For a more detailed description of the KPI-abbreviations, see Appendix 1 and 2.

Hypothesis 1 (Playing position). Six different playing positions exist which can be identified from event data. The different playing positions are listed in Table 1.

Table 1: Playing positions and their abbreviations.

Playing Position	Abbreviation
Striker	ST
Central midfielder	CM
Outer winger	OW
Full back	FB
Centre back	CB
Goalkeeper	GK

Hypothesis 2 (Playing style). For each of the existing playing positions, except the Goalkeeper position, different playing styles exist which can be identified from event data. They are listed, named and described in Tables 2 – 6.

Table 2: Hypothesis of existing playing roles for position ST-Striker.

1. ST - Striker				
Index:	Name:	Description:	Important KPIs:	Player examples:
1.1	The Powerstriker	<ul style="list-style-type: none"> - Old-school/traditional "heavy" player, good on deep through balls. - Functional technique. - Hard and dangerous shooting. - Good in the box. - Relatively fast, not necessarily quick. 	shots, tackles, headers, chall, tib, dze, prog carries	Haaland, Rooney, Blackstenius, Hegerberg, Schelin
1.2	The Poacher	<ul style="list-style-type: none"> - It's all about scoring goals, acting largely on that basis. - Good finisher. 	goals, xg, shots, tib, dze, sp xg, avg patch area	Lewandowski, Ronaldo, Gerd Müller, Hanna Ljungberg, Abby Wambach
1.3	The Artist	<ul style="list-style-type: none"> - Has an elegance that can not be missed. - Incredibly good technique. - Good assister and finisher, especially from difficult situations. 	goals, ass, xg, xa, xg share, dribbles, xp, plb, kp, tb, dze	Maradona, Bergkamp, Messi, Marta
1.4	The Worker	<ul style="list-style-type: none"> - Hard working team player. - Dueling strong. - Involved in a lot of situation not just close to opponent goal. - Good at setting up teammates. 	ass, xa, passes, crosses, headers, int, chall, wb oh, tb, avg patch area	Firmino, Pia Sundhage, Berg, Morata

Table 3: Hypothesis of existing playing roles for position CM-Central midfielder.

2. CM - Central midfielder				
<i>Index:</i>	Name:	Description:	Important KPIs:	Player examples:
2.1	The Box-to-box	<ul style="list-style-type: none"> - Strong running midfielder who moves over large parts of the pitch. - Can often be seen as a leading figure in the team. - "Does a bit of everything" (thus difficult to define important KPIs). 	tib, dze, avg patch area, gain	Kanté, Lampard, Amandine Henry
2.2	The Playmaker	<ul style="list-style-type: none"> - The player who controls the offensive game. - If the playmaker has a bad day, the team has made a bad day. - Big involvement in passing. 	pib, kp, tb, passes, crosses, gain, avg path area	Pirlo, Jorginho, Seger
2.3	The #10	<ul style="list-style-type: none"> - Moves in areas between the opponents' midfield and defense. - Decisive in the last third of the pitch (offensive) - Not always the most active in defense. 	pib, kp, tb, dribbles, directness, shots, assist, xA, gain	Zidane, Kaka, Dzsenerfer Marozsán, Asllani
2.4	The Anchor	<ul style="list-style-type: none"> - Good ball-winner. - The defending line's best friend. - A reverse trequartista. 	low directness and gain, tackles, challenges, fouls, xg share, gain, avg patch area	Gattuso, Vieira, Julie Ertz, Michelle Akers

Table 4: Hypothesis of existing playing roles for position OW-Outer winger.

3. OW - Outer winger				
<i>Index:</i>	Name:	Description:	Important KPIs:	Player examples:
3.1	The Solo-dribbler	<ul style="list-style-type: none"> - Challenges his/hers defender often. - Often lacking end product. - Somewhat selfish in actions. - Either favorite or hated by the audience. 	dze, dribbles, prog carries, low xP	Sterling, Delphine Cascarino, Jakobsson
3.2	The 4-4-2-fielder	<ul style="list-style-type: none"> - Strong in running. - Good foot. - Quick back to the "right side" of the pitch. - Somewhat of a coach favorite. 	crosses, directness, gain, challenges, tackles,	Sebastian Larsson, Beckham, Albrighton, Lombardo
3.3	The Star	<ul style="list-style-type: none"> - Both quick and fast. - Decisive in the last third. - Not always the most active in the defensive game. 	goals, xg, pib, kp, tb, dribbles, directness, shots, xg share	Salah, Rolfö/Schough

Table 5: Hypothesis of existing playing roles for position FB-Full back.

4. FB - Full back				
<i>Index:</i>	Name:	Description:	Important KPIs:	Player examples:
4.1	The Winger	<ul style="list-style-type: none"> - Usually involved in the attacking game. - Probably played as a central midfielder or winger as a junior. - Some limitations in the defensive game. 	crosses, gain, dribbles, dze, tib,	Marcelo, Trent, Ashley Lawrence
4.2	The Defensive-minded	<ul style="list-style-type: none"> - Good in defense. - Likes to relinquish responsibility in the offensive. - Limited in play with the ball. - Takes a lot of throw-ins. - Would probably be an okay central defender. 	tackles, challenges, fouls, headers, interceptions, low passing %	Kicki Bengtsson, Lustig, almost all Swedish full backs
4.3	The Box-to-box	<ul style="list-style-type: none"> - Good in defense. - Okay offensively. - Can come along in the attacks but is more limited offensively than Wingers. 	tackles, challenges, fouls, headers, interceptions, crosses, gain	Typical italian full back, ...

Table 6: Hypothesis of existing playing roles for position CB-Centre back.

5. CB - Centre back				
Index:	Name:	Description:	Important KPIs:	Player examples:
5.1	The Sweeper	<ul style="list-style-type: none"> - Very involved in the build up phase. - Good functional technology. - Likes to step up in the back of the attacking opponents. 	passes, lb, passes share, xP, gain	Nesta, Eriksson
5.2	The Leader	<ul style="list-style-type: none"> - Carries the team's defensive. - Good dueling game. - Not too bad with the ball. - Dangerous on set pieces. 	headers, tackles, challenges, interceptions, xg share, gain	Chiellini, Puyol, Fischer, Sjögran
5.3	The Physical	<ul style="list-style-type: none"> - Unpolished. - Likes to take a yellow. - Good physically. - Strong in the heading game. 	headers, tackles, challenges, lost balls, g mist, interceptions	Bailly, random english centre back

1.6 Disposition

Following the introduction, Section 2 is highly recommended for most readers to skip almost in its entirety. Here the theory for the developed models are described. Except Sections 2.3 – 2.4, which is about interesting football statistics, the theory is standard and uninteresting.

Section 3 is the main part of the report and starts off with going through and describes preconditions in the form of what there is to work with when it comes to data, where it originates from, its structure and what it contains. This information can be found in Section 3.1, but can easily be skipped by the not so detailed and football-data-interested-reader. However, Section 3.1.5 should still be read as it explains how we validate the results against the hypothesis. Then comes Section 3.2 and tells the story behind the different approaches towards the model development. After this rather unscientific descriptions of the models, they are explained in more detailed in the following Sections 3.3 – 3.5, which all are highly connected to Jupyter notebooks. Such notebooks are interactive computational environments/platforms useful for presenting data science projects [4]. Here it is used to display the models most important parts and results. These models are stored on GitHub, open for everyone to reach in a repository [5].

We recommend that readers, especially those with some fundamental programming knowledge, open up the notebooks and follow along simultaneous in the code while reading about the models in the report. Those with Git- and source-control-knowledge are also encourages to clone the Notebooks, go through the README, and try running the models with different settings than the default ones and see if they get any fun or useful results.

However, for those not so familiar with coding and notebook usage, we make sure to include some results for each of the models in Section 4. When it comes to the results presented only the results for the playing position 'ST', with its playing styles (see hypothesis in Section 1.5), are considered. Regarding the results for the rest of the playing positions some of them are found in the Appendix 2 – 3, and/or by alternating settings in the notebooks and running them.

Finally the models, results and findings are discussed in Section 5 and conclusions from the master thesis as a whole are documented in Section 6. The discussions and conclusions are mainly focusing on developed Model v.2, as it turned out to be the most interesting and useful model.

2 Theory

2.1 Normalization techniques

Normalization of a data set is essential for many machine learning estimators [6]. One commonly used normalization technique is to remove the mean and scaling to unit variance, we denote this normalisation as the standard score. For a set of observations/data points $\{x_i\}_{i=1}^n$, the standard score x'_i of an observation x_i can then be computed as

$$x'_i = \frac{x_i - \mu}{\sigma}, \quad (1)$$

with μ as the mean

$$\mu = \frac{1}{n} \sum_{i=1}^n (x_i) \quad (2)$$

and σ as the standard deviation

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}. \quad (3)$$

Another commonly used normalisation technique is by scaling the set of observations $\{x_i\}_{i=1}^n$ to a certain range, most often chosen to be $[0, 1]$, we denote this normalisation as the min-max score [7]. For a set of observations/data points $\{x_i\}_{i=1}^n$, the min-max score x'_i of an observation x_i , with feature range set to $[0, 1]$, can then be computed as

$$x'_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}. \quad (4)$$

2.2 Principal component analysis

Principal component analysis, or PCA for short, is a method used to rescale and, if desired, reduce dimensionality of data by projecting it into a linear subspace of the original data [6]. Given a set of original observations/data points in a matrix $\mathcal{T} = \{\mathbf{x}_i\}_{i=1}^n$, where \mathbf{x}_i is the i -th vector of length l , containing l number of statistics and n data vectors existing, i.e. $\mathcal{T} \in \mathbb{R}^{n \times l}$, we have the objective to learn a, possible lower-dimensional, representation $\mathbf{Z}_0 = \{\mathbf{x}_i^*\}_{i=1}^n$, with $\mathbf{Z}_0 \in \mathbb{R}^{n \times q}$ where $q \leq l$. Here \mathbf{x}_i^* is the i -th principle component and q the number of principle components. If q is chosen such that $q < l$ then the dimension is reduced and the remaining q principle components is a lower-dimensional representation of the original data.

How the alternative representation of the original data \mathcal{T} is learned, i.e. how the PCA model is learned, can be summaries down below in Method 1:

Method 1 Learn the PCA model

- 1: Compute the mean vector $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$
 - 2: Centre the data, $\mathbf{x}_{0,i} = \mathbf{x}_i - \bar{\mathbf{x}}$, for $i = 1, \dots, n$
 - 3: Construct the data matrix \mathbf{X}_0
 - 4: Perform SVD (Singular Value Decomposition) on \mathbf{X}_0 to obtain the factorization $\mathbf{X}_0 = \mathbf{U}\Sigma\mathbf{V}^\top$
 - 5: Compute the resulting principal components $\mathbf{Z}_0 = \mathbf{U}\Sigma = \mathbf{X}_0\mathbf{V}$
-

Here, \mathbf{V} corresponds to a change-of-basis in $\mathbb{R}^{n \times q}$ and is of the size $(l \times q)$. For further explanations of the steps in Method 1, especially steps 3 — 5 and SVD, see the book "*Supervised Machine Learning*" [6].

2.2.1 Scree plot

A scree plot displays how much of the variance of the original data \mathcal{T} is explained with each of the found principle components and can be used as a tool to decide on how many principle components q to keep (if PCA is used for dimension reduction) [8]. In order to create a scree plot the eigenvalues of the principle components are considered, since they represent the proportion of variance explained by each component, and ordered from highest value to lowest value.

2.2.2 PCA interpretation

When the PCA model has been learned, there are two matrices that are particularly interesting and that can be used to learn more about the transformation of the original data set \mathcal{T} . Firstly, since the columns of \mathbf{V} corresponds to the basis vectors of the new basis, the columns of \mathbf{V} contains information of how each of the l statistics are weighted in each principle component. The matrix \mathbf{V} can thus be used to understand the importance of the difference statistics in each principle component, and how certain statistics correlate to one another.

In \mathbf{Z}_0 we have the principal components \mathbf{x}_i^* , also known as the scores. Each row in \mathbf{Z}_0 gives an explanation of how each original vector \mathbf{x}_i scores in each principal component of the new basis $\mathbb{R}^{n \times q}$ [6].

2.3 Possession adjustment

Possession adjustment is a method to compute football statistics while taking the possession values of teams into account and is mostly used for defensive statistics. Adjusting the defensive statistics to the possession, i.e. as if the match was played with 50%/50% possession, gives further insight to the frequency of defensive actions [9].

Given a player statistic y_i , for example **tackles 90** (see Appendix A for further explanation), the possession adjusted statistic: **tackles 90 Padj**, is given by

$$y_i^{Padj} = y_i * \frac{50}{100 - P_i}, \quad (5)$$

where P_i is the possession for the team of player i .

2.4 Share KPIs

Share KPIs takes into account the team performance when computing some statistic of a player. Share KPIs can thus give a better understanding of a player's performance in respect to his or hers teammates.

Given a player statistic y_i , for example **tackles 90** (see Appendix A for further explanation), the share statistic: **tackles 90 share**, is given by

$$y_i^{share} = \frac{y_i}{\sum_{j=1}^m (y_j)}, \quad (6)$$

where m is the number of players in the team of player i .

2.5 Quantile classification

A quantile can be used to determine how many values in a distribution are above or below a certain (can be user defined) limit [10]. Given a set of observations $\{\mathbf{x}\}_i^n$, quantile classification determines whether each observation belong to the upper or lower group for a user chosen quantile q :

$$y_i(x_i, q) = \begin{cases} 1 & \text{if } x_i > x_q, \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

where x_q is the threshold value for the q :th quantile for the set of observations $\{\mathbf{x}\}_i^n$.

2.6 K-Means clustering

K-Means clustering aims to group a set of observations/training data points into K distinct clusters R_1, R_2, \dots, R_K , where each data point \mathbf{x}_i can only exist in exactly one cluster R_k . The clusters are found in such way that the distances to the cluster centers, summed over all data points $\{\mathbf{x}_i\}_{i=1}^n$, is minimized [6],

$$\arg \min_{R_1, R_2, \dots, R_K} \sum_{k=1}^K \sum_{\mathbf{x} \in R_k} \|\mathbf{x} - \hat{\boldsymbol{\mu}}_k\|_2^2. \quad (8)$$

Here $\hat{\mu}_k$ is the center of cluster R_k . That is, the mean of all data points $\{\mathbf{x}_i\}_{i=1}^n \in R_k$.

The minimization problem (8) is unfortunately a combinatorial problem and can thus not be solved exactly if the number of data points/observations n is large [6]. However, an approximate solution can be found by the following algorithm:

Algorithm 1 K-Means algorithm

```

1: Set the number of clusters K
2: Set the cluster centers  $\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_K$  to some initial values
3: while not converging do
4:   for  $i=1:1:n$  do
5:     Find closest cluster center  $\hat{\mu}_k$  to  $\mathbf{x}_i$ 
6:     Set  $\mathbf{x}_i$  to belong to cluster  $R_k$ 
7:   end for
8:   for  $k=1:1:K$  do
9:     Compute and update new cluster center  $\hat{\mu}_k$  as the average of all  $\mathbf{x}_i \in R_k$ 
10:  end for
11: end while

```

The algorithm has converged when the assignments of the cluster centers $\hat{\mu}_k$ no longer change.

2.7 Evaluation metrics

2.7.1 Baseline model

As good practice when working with a machine learning problem (or arguably other modelling problems) is to establish and/or create some sort of baseline model [6]. A baseline model is supposed to act as a reference for the performance level of the "real" model. In the simplest of cases a baseline model, for a classification problem, can be to just pick the most common class among the class labels in the training data and then use that for prediction. This is known as the Zero rate classifier; a classifier which classifies by always predicting the most frequent (or largest) class in the training data [11]. Another baseline model is the Random rate classifier. It applies prior knowledge from the training data by being a model that guesses at the weighted percentages of each of the existing class labels in the training data.

In this master thesis we also introduce and use an "simple" model which is considered to be a bit broader than just a baseline to beat with the "real" model. Here we instead refer to an "simple" model as a model solving a simplification of the original problem (classifying a football player's playing style). This model is then compared with the other, more advanced model, in order to put the hypothesis to test.

2.7.2 Confusion matrix

For evaluation and performance measuring of models used for classification problems a so called Confusion matrix can be used [6]. It is a simple matrix, which can be plotted in the form of a heat-map or just with values written out inside of it. It displays all the True negative (TN), False negative (FN), False positive (FP) and True positive (TP) values of the validation data. For a multi-class classification model, lets say a 3-class classification model with Class A, Class B and Class C, we then have for Class A:

- **TN** - values predicted as not Class A from the model which actually are not Class A.
- **FN** - values predicted as Class A from the model which actually are not Class A.
- **FP** - values predicted as not Class A from the model which actually are Class A.
- **TP** - values predicted as Class A from the model which actually are Class A.

Having, from validation data, y (actual output values), $\hat{y}(\mathbf{x})$ (model predicted values) and possible distinct output classes $y \in \{A, B, C\}$ the confusion matrix, with Class A as reference, would be constructed as:

Table 7: Confusion matrix template.

	$y = A$	$y = B$	$y = C$
$\hat{y}(\mathbf{x}) = A$	TP	FP	FP
$\hat{y}(\mathbf{x}) = B$	FN	TN	TN
$\hat{y}(\mathbf{x}) = C$	FN	TN	TN

2.7.3 Precision, Recall, Specificity and Accuracy

Precision, or positive predictive value (PPV), shows what fraction of predictions as a positive class were correctly classified as positive:

$$\mathbf{Precision} = PPV = \frac{TP}{TP + FP}. \quad (9)$$

Recall, also referred to as True Positive Rate (TPR), shows what fraction of the positive predictions that were correctly classified as positive:

$$\mathbf{Recall} = TPR = \frac{TP}{TP + FN}. \quad (10)$$

Specificity, also referred to as True Negative Rate (TNR), shows what fraction of the negative predictions that were correctly classified as negative:

$$\mathbf{Specificity} = TNR = \frac{TN}{TN + FP}. \quad (11)$$

Lastly, accuracy shows the fraction of the total number of predictions that were correctly classified:

$$\mathbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \quad (12)$$

3 Method and Model development

3.1 Data background

The data used in this thesis has in its entirety been provided by PlayMaker.AI or created by us. It consists of five different types of data sources/types. There is the data in its rawest form, match event data. There is processed data, computed from the match event data: KPI player statistics data, team performance data and playing position data. Playing position data also exist in the form of scraped data. Then there is the data created in this master thesis with the sole purpose to validate the developed models, namely the validation data set. Down below in Sections 3.1.1 – 3.1.5 the data sources/types are described in more detail.

In this master thesis data from two seasons are used: Allsvenskan 2021 and Premier League 2021/2022 (up until 2022-02-23). The two leagues have been chosen because those are probably the ones that people in Sweden follow the most, and also, due to the fact that PlayMaker mostly works with clubs in Sweden. The reason why seasonal data, i.e. data from several matches in a season over longer periods of time, is used has to do with PlayMaker.AI (the platform) being built up that way and how it in a good way represents a player's statistical playing contribution.

3.1.1 Match event data

Match event data contain events that have occurred during a match in which the ball is involved. Examples of such match events are passes, shots, interceptions, etc. For each type of event it is labelled which player it belongs to, timestamp of when during the match it happens and in which position, in the form x- and y-coordinates, of where on the pitch it happens. For each match approximately 1700 to 2300 observations of events are yielded with the above mentioned labelling. From these further statistics are calculated, both for individual players and teams as a whole, leading to the KPI and team performance data.

There are also two other types of data which can be collected during a match, namely GPS-data and tracking data. GPS-data is as it sounds is data from GPS devices worn by the players during the match, collecting their positions, movements, velocities and accelerations. However, such data is mainly collected and owned by the teams themselves, used for physical measurements and checkups. Tracking data is collected via special cameras following the ball and players on the pitch for every single moment of the game. Since PlayMaker does not yet work with neither GPS- nor tracking data they are disregarded in this master thesis.

PlayMaker collects match event data themselves utilizing so called tagging; a match is played back and events are tagged in terms of events as described above. PlayMaker also buys match event data from other companies collecting it and parse it in order to get it into the same format as their own data.

A detailed and comprehensive description of how match event data is collected is given in the article *"A public data set of spatiotemporal match events in soccer competitions"* [12].

3.1.2 KPI data

As mentioned in the section above the KPI data is based on, and computed from, the match event data. Although KPIs can be computed and used for a group of players or teams as a whole, if nothing else is stated, referring to player-KPIs throughout this master thesis. In other words, a KPI is a "Key Performance Indicator" of an individual football player's performance on the pitch.

Some of the KPIs are straightforward and rather self-explanatory, especially for someone with common football knowledge. Other KPIs are more complicated, both in how they are computed and in what they are indicating in terms of performance. All KPIs utilized in this master thesis, with a brief explanation, are listed down in Appendix 1 – 2. The KPIs can be divided into two different types: `KPI_org` - KPIs already existing on PlayMaker.AI platform which can be used right away and `KPI_new` - new KPIs computed.

3.1.3 Team performance data

The team performance data is very much alike the KPI data in the sense of it mainly being computed from match event data and contains information regarding team performance over time. It is available on PlayMaker.AI and can be used directly.

The team performance data used in this master thesis are ball possession values, i.e. indicators of how large percentages each team is in possession (control) of the ball.

3.1.4 Playing position data

Playing position data refers to the data which contain information about the playing position a certain player belongs to, and plays in. There exists data on PlayMaker.AI for playing positions of two sorts. The first one being by default scraped from online football team websites, but can also be manually set by users of the platform. The other one being a computed playing position, based on estimates from the match event data.

3.1.5 Validation data set

The arguably most important data for this master thesis is the validation data set. Its importance emerges from Assumption 5, that it is considered to contain the truth in terms of the results this master thesis aim to achieve. As already explained in Section 1.4 this is a big limitation. For reasons which becomes clearer in Section 3.4 (particularly in Section 3.5.4) two different validation data sets were created in this master thesis. In terms of design they are structured in the same way; containing data in the form of football players from Premier League and Allsvenskan labelled after playing positions and playing styles. The labelling, essentially the creation of the validation data set, was done manually by going through all available football players for the considered seasons (see above) and determining each player's "true" playing style, and position. This was done in a subjective manner by being based on common football knowledge of the master thesis participants in close dialog with supervisors at PlayMaker. The validation data set was cross-validated between the participants and, as already mentioned, revised later on in the master thesis.

Problems and possibilities in regards to using a validation data set such as this one are debated mainly in Section 5.3, but also in the conclusions, Section 6. As of now we justify the choice with simply stating that football in itself is very much subjective.

The original validation data set, which has the players labelled after the hypothesis in Section 1.5, has the distribution between the different playing styles as seen down below in Figure 1.

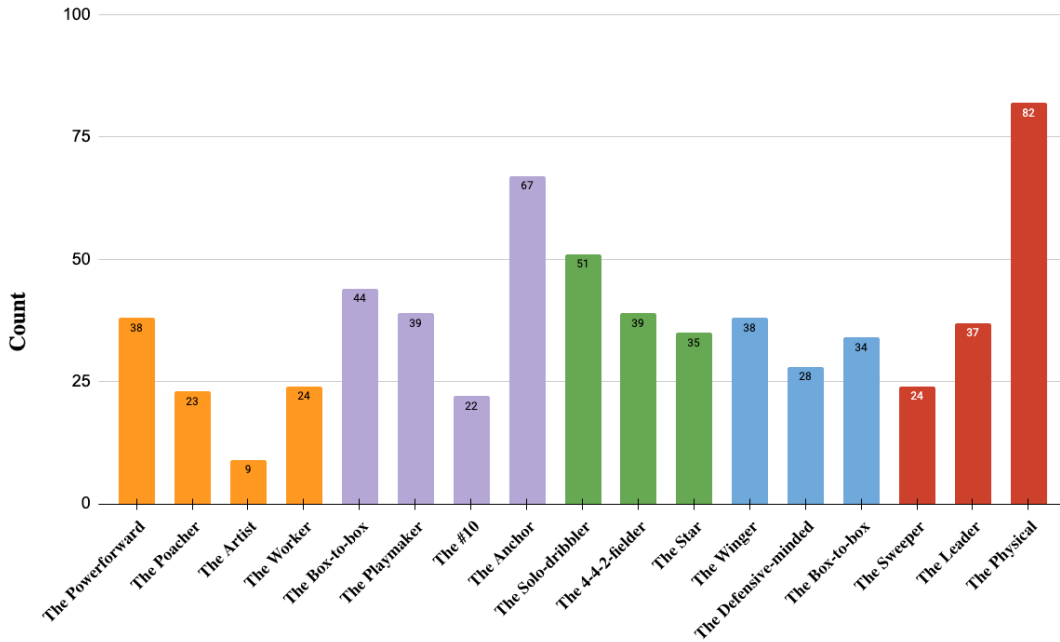


Figure 1: Count over the number of different players in each playing style for the original validation data set. In total there are 634 unique players.

3.2 Modelling approach

The approach towards developing and creating a model which fulfil the aim and goals specified in Section 1.2 starts with solving the problem of finding the player's playing positions, as defined in Definition 2 and suggested in Hypothesis 1. Since the existing playing position data had a low accuracy against our validation data set we early on, in discussions with PlayMaker, decided that a new position detection model had to be developed.

The performances of the two existing playing position data for both Premier League 2021/2022 and Allsvenskan 2021 can be seen in Table 8 down below.

Table 8: Performance accuracies of the two existing playing position data versus baseline models accuracies.

Model	PL 2021/22	Allsvenskan 2021
Zero rate	0.27	0.27
Random rate	0.21	0.21
PlayMaker estimated	0.58	0.56
PlayMaker primary	0.74	0.66

See Section 2.7.1 for how Zero rate and Random rate were computed. The new developed Position detection model is presented and described in a similar way to the existing ones in Section 3.3 down below. The new results, see Section 4.1 for them, turned out to be a significant improvement; hence it was an easy decision to use the new model throughout the rest of the master thesis. In Section 5.1 the Position detection models are discussed more in-depth and thoughts about why they might not be as important in the future are explained.

Once the problem of finding the player's positions had been sorted out to an acceptable level the focus shifted towards the main target: to develop a model which classifies an individual football player's playing style, as defined in Definition 1. The first approach towards this was to start off simple, with a "simple" model that essentially acts as a reference as explained in Section 2.7.1. The "easiest" way of categorising a footballer's playing style was thought to be in a binary way, where a player's playing style either is "Offensive" or "Defensive". The names explain the meaning of the playing styles well with their straightforward definitions: An offensive player prioritizes attacking and is more involved in offensive play. A defensive player prioritizes defending and is more involved in defensive play. The "simple" model developed for this is presented in Section 3.4 with results in Section 4.2. It is referred to as Model v.1 and used together with the Position detection model in order to find the different playing styles for each playing position.

With a functional "simple" model we could start working against developing a model which would identify and classify playing styles after Hypothesis 2 in Section 1.5. The model approach used was an unsupervised PCA-based model. By letting the data speak for itself and by interpreting the emerging results we could find out whether or not the playing styles in the hypothesis did in fact exist and which players were classified within the respective found playing styles. The developed model is in its entirety presented in Section 3.5, its results in Section 4.3, and referred to as Model v.2. It is also position dependent and used together with the Position detection model for finding playing styles in different positions, as the hypothesis suggests.

3.3 Position detection model

This section presents and describes the model used for detecting playing positions of football players according to the definition given in Definition 2. The model is based on clustering techniques and evaluated against the validation set created and built up according to Section 3.1.5. The finalised and implemented Position detection model is later on used in the preprocessing stages for Model v.1 and Model v.2.

The main idea behind the Position detection model is to cluster all players into six, since we assume there are six existing playing positions, different groups representing the playing positions in Table 1. The clustering is carried by a combination of where the players are doing most of their passes on the pitch, in terms of the x- and y-coordinates of the pass, together with other KPIs which helps with separating the playing positions.

A simplified implementation of the model, where some of the steps described in Section 3.3.2 are hidden in modules, can be found in the Jupyter Notebook `position_detection.ipynb` [13]. Please feel free to follow along in the code.

3.3.1 Problem/Mathematical formulation

Given a football player u with seasonal KPI data and match event data we want to detect its playing position pos , where $pos \in \{ST, CM, OW, FB, CB, GK\}$ according to Table 1. This can be formulated as a clustering classification problem where we want to cluster all available football players \mathbf{u} from the data into six different clusters.

3.3.2 Scheme and description

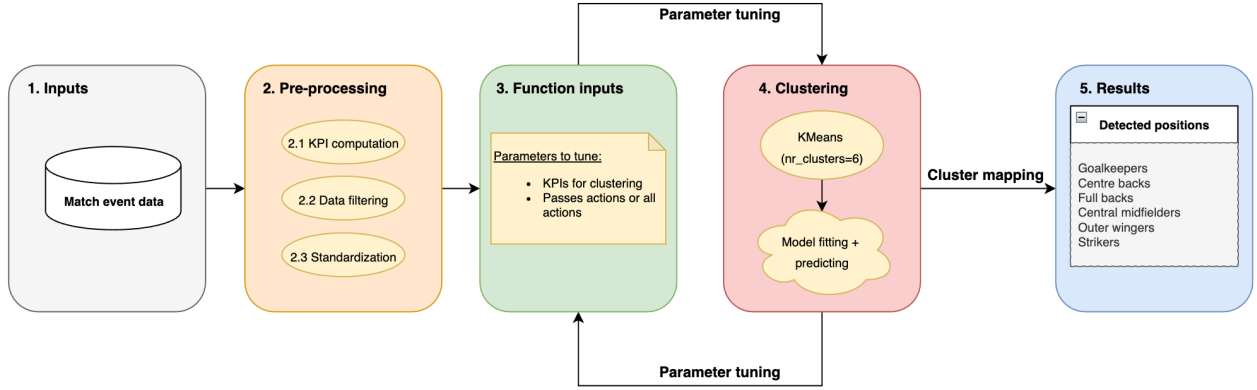


Figure 2: Scheme over position detection model.

The model implementation builds upon five main steps, as can be seen above in Figure 2. Those steps are, and does, the following:

1. **Inputs** - As inputs to the model there is the match event data, described in Section 3.1.1. The data needs to exist for all players of which positions are to be detected for.
2. **Pre-processing** - Here the inputted data is pre-processed to work with the model in terms of: more KPIs being computed, see Section 3.1.2, filtering out players which have played less than 360 minutes from the match event data and data standardized according to Equation (4).
3. **Function inputs** - As inputs to the function doing the detection there is the KPIs used for the clustering and if only passing actions should be considered. Choosing which KPIs to include when clustering and whether or not to just look at where the passing actions occur on the pitch is done iterative in the parameter tuning process. In Section 3.3.3 the model tuning process is explained.
4. **Clustering** - This step clusters the players using K-Means, see Section 2.6 and Equation (8), with number of clusters set to six via fitting of the model and predicting cluster belonging for each player.
5. **Results** - From mapping of the found clusters to actual playing position, see function `map_clusters_to_position()` [14], the results are obtained. In Section 4.1 the results are presented.

3.3.3 Model tuning

The model tuning process consist of choosing which KPIs, together with the average passing coordinates, to include to better separate players between positions. This tuning have been done in an iterative manner and the final KPIs used for the results are the following: [xg 90, dribbles 90, tib 90, dze 90, chall %, int 90, dribb past 90, passes share, gain 90, headers 90].

3.4 Model v.1

In this section Model v.1 is presented together with its results. This model has been developed to act as a "simple" model. The resulting performance of Model v.1 will then be aimed to be overcome by the performances of the other, more advanced and detailed, developed model.

A simplified implementation of the model, where some of the steps described in Section 3.4.2 are hidden in modules, can be found in the Jupyter Notebook `model_v1.ipynb` [15]. Please feel free to follow along in the code.

3.4.1 Problem/Mathematical formulation

Given a football player u with match event data and team possession data we want to detect its playing style $y(pos_u)$, where $y(pos_u) \in \{Offensive, Defensive\}$, for the detected playing position pos_u . This can be formulated as a binary classification problem where we want to classify all available football players \mathbf{u} from the data into the two groups $\{Offensive, Defensive\}$. The classification is also position dependent, therefore each position $pos \in \{ST, CM, OW, FB, CB\}$ (goalkeeper group neglected as explained in Section 1.4) should have its own classifier within that position group.

3.4.2 Scheme and description

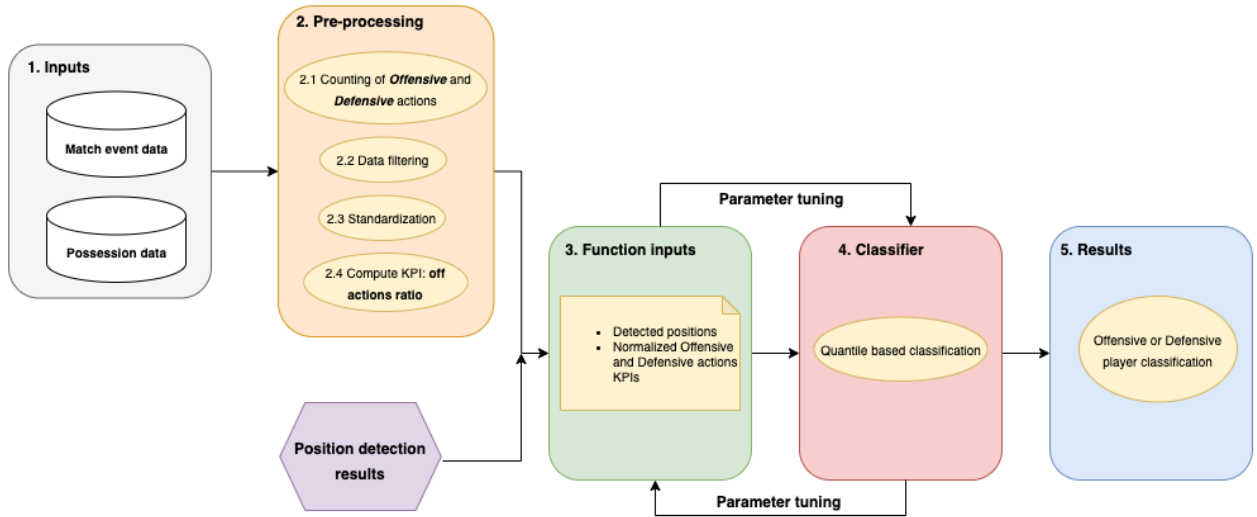


Figure 3: Scheme over playing style Model v.1.

The Model v.1 implementation can be seen above in Figure 3. The five steps of the model does the following:

1. **Inputs** - As inputs to the model there is the match event data, described in Section 3.1.1, and possession data described in Section 3.1.3.
2. **Pre-processing** - The pre-processing part of model 1 has 4 main steps. The first step, 2.1, counts the number of offensive and defensive actions a player has made from the match event data input. The actions that counts as offensive and defensive are listed in Appendix B. When the counting of actions is done, the model filter out players which have played less than 360 minutes from the match event data. In step 2.3 the number of actions are normalized by per 90 minutes played and then lastly the number of defensive actions are normalized by possession of the team the player plays for. Finally, the model computes the KPI: offensive actions ratio (see Appendix 2 for further KPI explanations).
3. **Function inputs** - Inputs to the Model v.1 classifier are the detected positions from the position detection model and the offensive actions ratio KPI r_{off} , for each player $u \in \mathbf{u}$.
4. **Classifier** - Each player gets classified as either an "Offensive" or a "Defensive" player according to Equation (7) with r_{off} as observations \mathbf{X} , and with the quantile input argument q as position dependent, i.e. $q = q(pos)$. Thus, the classifier can be tuned with the parameter $q(pos)$ for each position independently.
5. **Results** - The results from the explained classifier is whether the player is considered an "Offensive" or "Defensive" player in his or her playing position. In Section 4.2 the results are presented.

3.4.3 Model tuning

The tuning process of Model v.1 consist of finding "suitable" values for $q(pos)$. This was done in an iterative manner and the final values for each position are presented in Table 9 below.

Table 9: Final quantile settings for all positions.

Position	q
'ST'	0.25
'CM'	0.25
'OW'	0.30
'FB'	0.40
'CB'	0.70

3.4.4 Playing style to "Offensive"/"Defensive"-classification mapping

The results presented in the next section have been extracted by comparing the model results to the validation data set, described in Section 3.1.5. Since the players in the validation set have been classified with playing styles, and not as either "Offensive" or "Defensive", we need to map the playing styles to "Offensive"/"Defensive"-classification to be able to compare the validation set with the model results. Below in Table 10 this mapping can be seen.

Table 10: Mapping of playing styles to "Offensive"/"Defensive"-classification.

Playing style index	Playing style	Offensive/Defensive
1.1	The Powerforward	Offensive
1.2	The Poacher	Offensive
1.3	The Artist	Offensive
1.4	The Worker	Defensive
2.1	The Box-to-box	Offensive
2.2	The Playmaker	Defensive
2.3	The #10	Offensive
2.4	The Anchor	Defensive
3.1	The Solo-dribbler	Offensive
3.2	The 4-4-2-fielder	Defensive
3.3	The Star	Offensive
4.1	The Winger	Offensive
4.2	The Defensive-minded	Defensive
4.3	The Box-to-box	Offensive
5.1	The Sweeper	Offensive
5.2	The Leader	Defensive
5.3	The Physical	Defensive

3.5 Model v.2

This section presents Model v.2 together with its results. Model v.2 looks further than only categorising players into the two groups $\{Offensive, Defensive\}$. With the hypotheses of playing styles presented in Section 1.5 in mind, this model aims to look at what the available and unlabelled data actually tell us about what playing styles that might exist. Model v.2 is thus an unsupervised model. There is also nothing that strictly decides that the playing styles explained in the hypothesis must be the final playing styles detected by the unsupervised model.

A simplified implementation of the model, where some of the steps described in Section 3.5.2 are hidden in modules, can be found in the Jupyter Notebook `model_v2.ipynb` [16]. Please feel free to follow along in the code.

3.5.1 Problem/Mathematical formulation

Given a football player u with match event data, KPI data and team possession data we want to detect its playing style $y(pos_u)$, where $y(pos_u) \in \{y_1(pos), \dots, y_{n(pos)}(pos)\}$ with $pos \in \{ST, CM, OW, FB, CB\}$. Here $n(pos)$ is the number of playing styles for position pos .

Together with finding the playing style for player u , we also want the model to be able to tell us how much player u correlates to each found playing style for each position. Thus, given a football player u with match event data, KPI data and team possession data, we want to detect its playing style correlation vector \mathbf{y}_u , with vector length of $\sum_{pos \in \{ST, CM, OW, FB, CB\}} n(pos)$.

Combining the two above explained problem formulations, we have a regression problem where we want to find the playing style correlation vector \mathbf{y}_u for all available players \mathbf{u} . For each player u , the playing style $y(pos_u)$ can then be found from the highest correlation weight in \mathbf{y}_u belonging to position pos_u .

3.5.2 Scheme and description

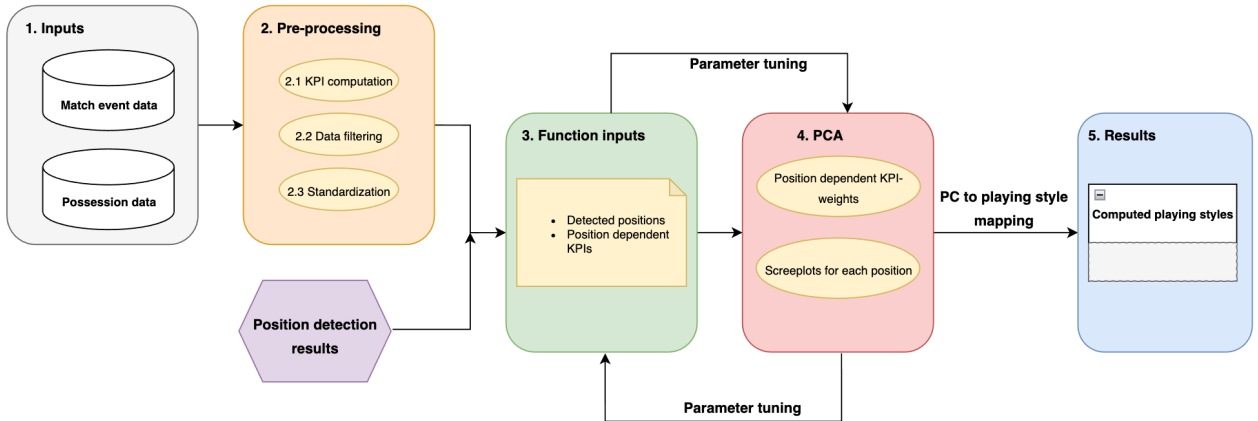


Figure 4: Scheme over playing style Model v.2.

The model implementation for Model v.2 builds upon five main steps, as can be seen above in Figure 4. Those steps are, and does, the following:

1. **Inputs** - As inputs to the model there is the match event data, described in Section 3.1.1, and possession data from Team performance data described in Section 3.1.3.
2. **Pre-processing** - Here the inputted data is pre-processed to work with the model in terms of: more KPIs being computed, see Section 3.1.2, filtering out players which have played less than 360 minutes from the match event data and data standardized according to Equation (1).
3. **Function inputs** - Inputs to the function are all the players u with their detected positions and position dependent KPIs. The choice of which KPIs to use for each position is done iteratively and explained in the parameter tuning process, Section 3.5.3.
4. **PCA** - The PCA-step consist of looking at the principal component weight patterns and scree plots with the goal of finding playing styles within the data. This is done using `plot_PCA_screepplot()` and

`plot_PCA_weights()` [17]. The found playing styles should both be reflected by the weight patterns and also be recognizable by an "ordinary" football enthusiast (see also Section 3.5.3 for more details).

5. **Results** - Following the PCA-step the model maps the principal components to the found playing styles (see `dict_playingStyle_example_mapper` in [16]). After the mapping, the model results are the playing style $y(pos_u)$ and the playing style correlation vector \mathbf{y}_u for all players $u \in \mathbf{u}$. The correlation vector \mathbf{y}_u is computed from the resulting PCA-models weight matrices \mathbf{V}_{pos} . Note that there exists a weight matrix \mathbf{V} for each position pos which contains the weights from the found playing styles in position pos . For further details about the computation of \mathbf{y}_u see `models_lib.py` [18].

3.5.3 Model tuning

The tuning process for Model v.2 consists mainly about finding suitable KPI settings/configurations for the different playing positions. This was done in an iterative manner by tracking changes to the scree plots and PCA weights for the computed PCs (principle components) for each position. The aim should be to have as much variance of the data as possible explained by PCs that in some way could represent a playing style (either which already exists in our hypothesis or one that the user believes should exist).

For further details and insights into the tuning process, and to see which settings were used to get results in this reports, see `config.py` [19].

3.5.4 Revised/found playing styles

When the tuning of the model is completed, a certain number of playing styles for each position have been found. These playing styles do not necessarily need to correlate exactly to the Hypothesis 2 in Section 1.5. Thus, to evaluate the performance of Model v.2, a revised validation data set with labelling of playing styles following the found model playing styles had to be created. The revised playing styles, together with the performance measurements, can be found in Section 4.3.

4 Results

4.1 Position detection model

Below are the results for position detection model presented both for Premier League 2021/2022 season and Allsvenskan 2021 season in Table 11, and Figures 5 and 6. For more results in the form of confusion matrices and class metric for each playing position see Appendix 1. The evaluation metrics have been computed by comparing the model result to the positions in the validation data set, see Section 3.1.5.

Table 11: Performance accuracies of the two existing playing position data and baseline models accuracies versus the new developed position detection model.

Model	PL 2021/22	Allsvenskan 2021
Zero rate	0.27	0.27
Random rate	0.21	0.21
PlayMaker estimated	0.58	0.56
PlayMaker primary	0.74	0.66
New Model (pre tuning)	0.71	0.55
New Model (post tuning)	0.89	0.84

In Table 11 (pre tuning) is the case when only clustering for the coordinates of the passes are considered and (post tuning) includes also the KPIs explained in Section 3.3.3. For explanation of the baseline models accuracies, see Section 2.7.1.

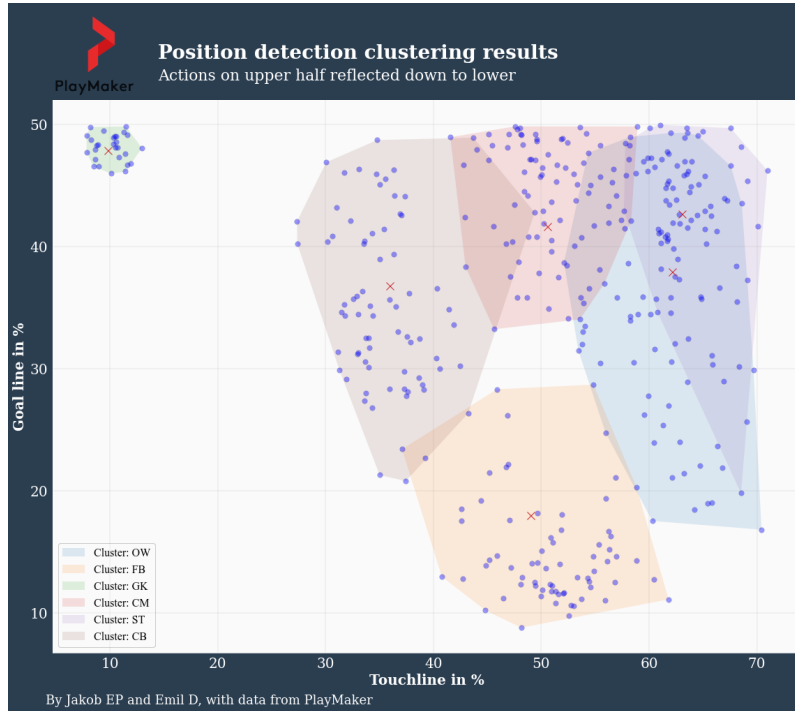


Figure 5: Resulting found clusters from position detection model for Premier League 2021/2022 season. The clusters are plotted on half a pitch, average pass positions reflected down to lower (see Assumption 9).

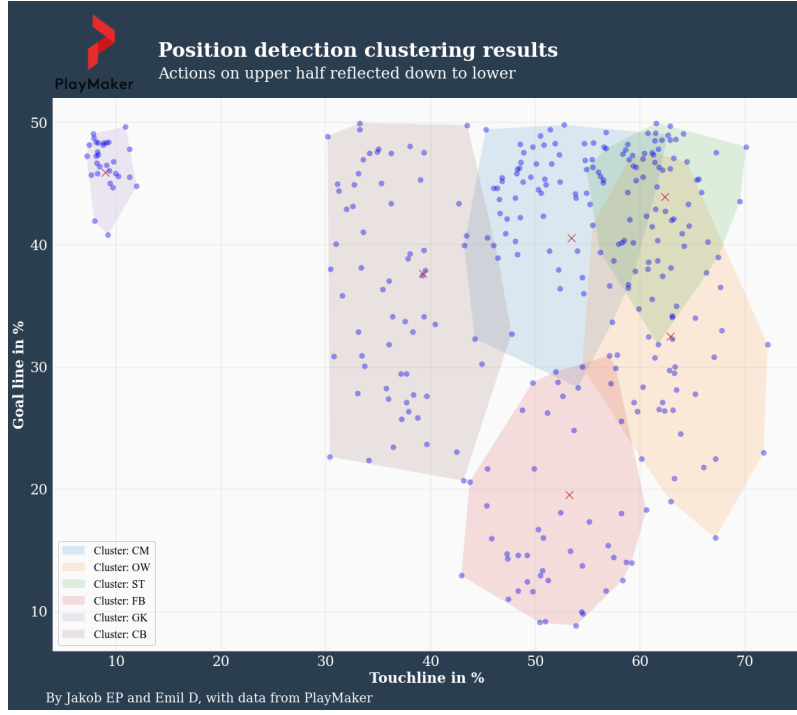


Figure 6: Resulting found clusters from position detection model for Allsvenskan 2021 season. The clusters are plotted on half a pitch, average pass positions reflected down to lower (see Assumption 9).

4.2 Model v.1

Below are the results for Model v.1 presented. The confusion matrices, together with their class metrics (computed by comparing the model result to the validation data set), can be found in Appendix 2. Mapping of playing styles have been done according to Table 10 and used model parameters according to Table 9.

Table 12: Performance accuracies for playing style classifications per position of Model v.1 versus baseline models accuracies (explained in Section 2.7.1).

Position	Accuracies		
	Zero rate	Random rate	Model v.1
'ST'	0.78	0.67	0.67
'CM'	0.73	0.61	0.50
'OW'	0.71	0.59	0.65
'FB'	0.61	0.53	0.65
'CB'	0.73	0.61	0.64

4.3 Model v.2

In Sections 4.3.1 — 4.3.4 the results from Model v.2 are presented in terms of scree plot, playing style mapping example, the found playing styles, performance against revised validation data set and player examples from both Allsvenskan and Premier League.

4.3.1 Mapping results example

From Figures 7 and 8 down below the playing style "The Target", for position 'ST', was found. For complete mapping results see dict_playingStyle_example_mapper in [16].

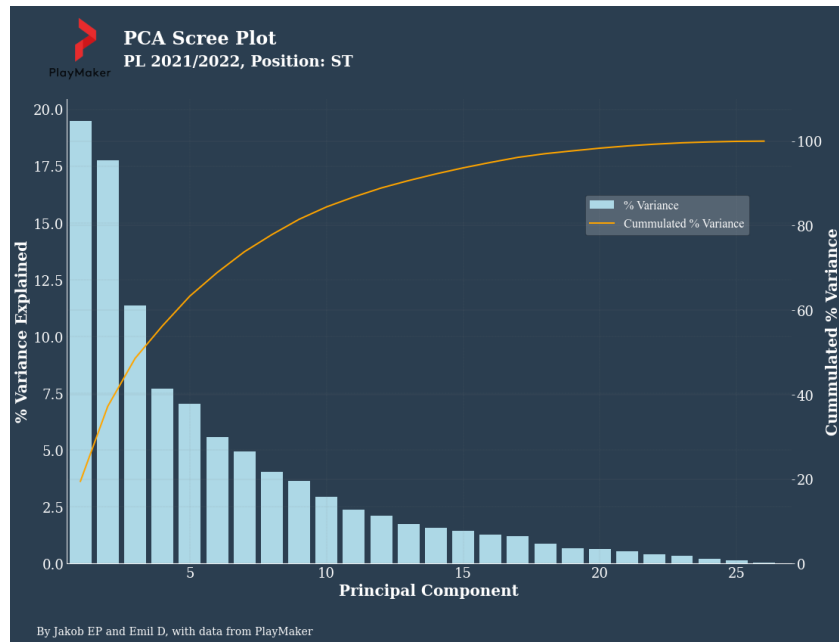


Figure 7: Resulting scree plot from PCA, position 'ST'.

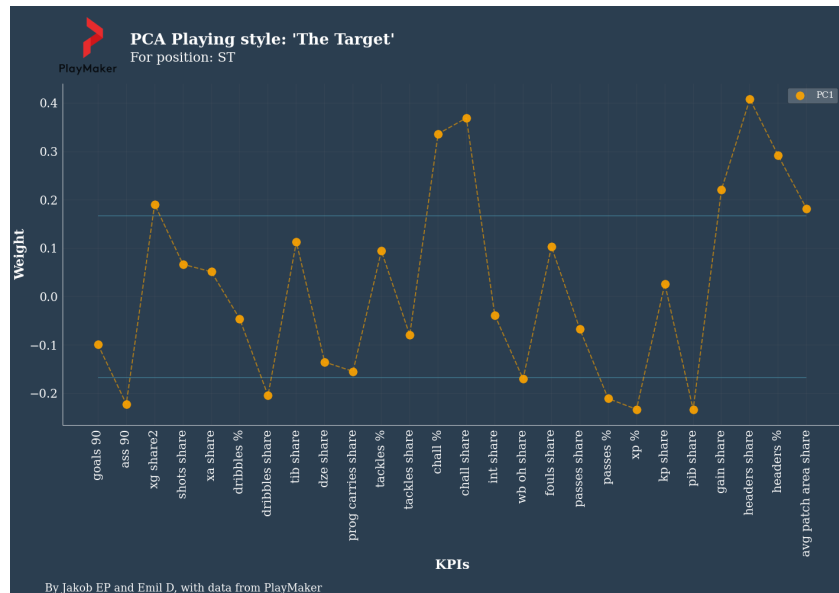


Figure 8: Resulting PCA weights for KPIs belonging to PC1, position 'ST'. Mapped to playing style "The Target".

4.3.2 Found playing styles

Below the found playing styles from Model v.2, after the result mapping had been carried out (for all positions, not just 'ST'), are listed, named and described in Tables 13 – 17.

Table 13: Found playing styles from Model v.2 for position ST-Striker.

1. ST - Striker				
<i>Index:</i>	Name:	Description:	Important KPIs:	Player examples:
1.1	The Target	<ul style="list-style-type: none"> - Bad at dribbling and carrying the ball. - Good with the head. - Often tall. - Good in the box. - Perhaps a bit slow. 	shots, tackles, headers, chall, tib, dze	Peter Crouch
1.2	The Poacher	<ul style="list-style-type: none"> - It's all about scoring goals, acting largely on that basis. - Good finisher. 	goals, xg, shots, tib, dze, sp xg, avg patch area	Lewandowski, Ronaldo, Gerd Müller, Hanna Ljungberg, Abby Wambach
1.3	The Artist	<ul style="list-style-type: none"> - Has an elegance that can not be missed. - Incredibly good technique. - Good assister and finisher, especially from difficult situations. 	goals, ass, xg, xa, xg share, dribbles, xp, pib, kp, tb, dze	Maradona, Bergkamp, Messi, Marta
1.4	The Worker	<ul style="list-style-type: none"> - Hard working team player. - Dueling strong. - Involved in a lot of situation not just close to opponent goal. - Good at setting up teammates. 	ass, xa, passes, crosses, headers, int, chall, wb oh, tb, avg patch area	Firmino, Pia Sundhage, Berg, Morata

Table 14: Found playing styles from Model v.2 for position CM-Central midfielder.

2. CM - Central midfielder				
<i>Index:</i>	Name:	Description:	Important KPIs:	Player examples:
2.1	The Box-to-box	<ul style="list-style-type: none"> - Strong running midfielder who moves over large parts of the pitch. - Can often be seen as a leading figure in the team. - "Does a bit of everything" (thus difficult to define important KPIs). 	tib, dze, avg patch area, gain	Kanté, Lampard, Amandine Henry
2.2	The Playmaker	<ul style="list-style-type: none"> - The player who controls the offensive game. - Big involvement in passing. - Both the #10 and more deep-lying playmaker in this hypothesis. 	pib, kp, tb, passes, crosses, gain, avg path area	Pirlo, Jorginho, Seger
2.3	The Anchor	<ul style="list-style-type: none"> - Good ball-winner. - The defending line's best friend. 	low directness and gain, tackles, challenges, fouls, avg patch area	Gattuso, Vieira, Julie Ertz, Michelle Akers

Table 15: Found playing styles from Model v.2 for position OW-Outer winger.

3. OW - Outer winger				
<i>Index:</i>	Name:	Description:	Important KPIs:	Player examples:
3.1	The Solo-dribbler	<ul style="list-style-type: none"> - Challenges his/hers defender often. - Often lacking end product. - Somewhat selfish in actions. - Either favorite or hated by the audience. 	dze, dribbles, prog carries, low xP	Sterling, Delphine Cascarino, Jakobsson
3.2	The 4-4-2-fielder	<ul style="list-style-type: none"> - Strong in running. - Good foot. - Quick back to the "right side" of the pitch. - Somewhat of a coach favorite. 	crosses, directness, gain, challenges, tackles,	Sebastian Larsson, Beckham, Albrighton, Lombardo
3.3	The Star	<ul style="list-style-type: none"> - Both quick and fast. - Decisive in the last third. - Not always the most active in the defensive game. 	goals, xg, pib, kp, tb, dribbles, directness, shots, xg share	Salah, Rolfö, Schough

Table 16: Found playing styles from Model v.2 for position FB-Full back.

4. FB - Full back				
Index:	Name:	Description:	Important KPIs:	Player examples:
4.1	The Winger	<ul style="list-style-type: none"> - Usually involved in the attacking game. - Probably played as a winger as a junior. - Some limitations in the defensive game. 	crosses, gain, dribbles, dze, tib, pib, prog carries	Robertsson, Ashley Lawrence
4.2	The Defensive-minded	<ul style="list-style-type: none"> - Good in defense. - Likes to relinquish responsibility in the offensive. - Limited in play with the ball. - Takes a lot of throw-ins. - Would probably be an okey central defender. 	tackles, challenges, fouls, headers, interceptions, low passing %	Kicki Bengtsson, Lustig, almost all Swedish full backs
4.3	The Inverted	<ul style="list-style-type: none"> - Very involved in the offensive build up. - Probably played as a central midfielder as a junior. - A good candidate to deliver the pieces. 	crosses, passes share, gain, pib	Cancelo, Trent Alexander Arnold

Table 17: Found playing styles from Model v.2 for position CB-Centre back.

5. CB - Centre back				
Index:	Name:	Description:	Important KPIs:	Player examples:
5.1	The Leader	<ul style="list-style-type: none"> - Carries the team's defensive. - Good dueling game. - Also good with the ball. - All around good! 	headers, tackles, challenges, interceptions, gain, passes share	Chiellini, Puyol, Fischer, Sjögran
5.2	The Low-risk-taker	<ul style="list-style-type: none"> - Involved in offensive build up. - Safe passing though. - Perhaps not the strongest defensively. 	xP %, passes share, passes %	
5.3	The Physical	<ul style="list-style-type: none"> - Unpolished. - Likes to take a yellow. - Good physically. - Strong in the heading game. 	headers, tackles, challenges, lost balls, g mist, interceptions	Bailly, random english centre back

From the found playing styles, a revised validation data set was created which is summarised in Figure 9 down below.

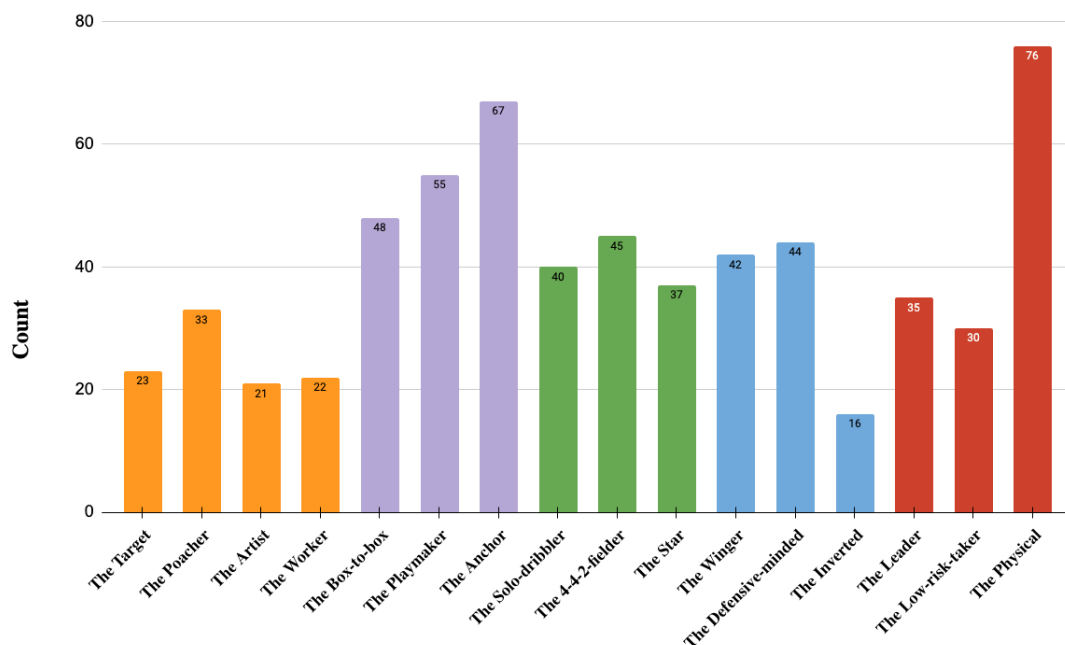


Figure 9: Count over the number of different players in each playing style for the revised validation data set. In total there are 634 unique players.

4.3.3 Performance

Below are the performance results for Model v.2 presented with the confusion matrix together with class metrics for the position 'ST'. The presented results are from both Premier League 2021/2022 season and Allsvenskan 2021 season combined but normalized separately by Equation (1) due to differences in league-level. The confusion matrices and class metrics for the other positions can be found in Appendix 3.

Table 18: Performance accuracies for playing style classifications per position of Model v.2 versus baseline models accuracies (explained in Section 2.7.1).

Position	Accuracies		
	Zero rate	Random rate	Model v.2
'ST'	0.34	0.26	0.79
'CM'	0.46	0.36	0.64
'OW'	0.35	0.33	0.70
'FB'	0.43	0.38	0.74
'CB'	0.53	0.39	0.70

Table 19: Confusion matrix and belonging class metrics Model v.2 for position 'ST'.

	1.1	1.2	1.3	1.4	nr of actual	precision	recall	specificity	F1-score
1.1	20	0	2	1	23	0.87	0.74	0.95	0.80
1.2	4	22	4	1	31	0.71	0.96	0.87	0.81
1.3	0	1	17	0	18	0.94	0.65	0.98	0.77
1.4	3	0	3	13	19	0.68	0.87	0.92	0.76
nr of predicted	27	23	26	15	91				

4.3.4 Player examples

Examples of four different players, with their assigned/classified playing styles and correlation to some other playing styles are presented in Spider plots in Figures 10 and 11 down below. The players in Figure 10 plays in Allsvenskan and those in Figure 11 plays in Premier League.

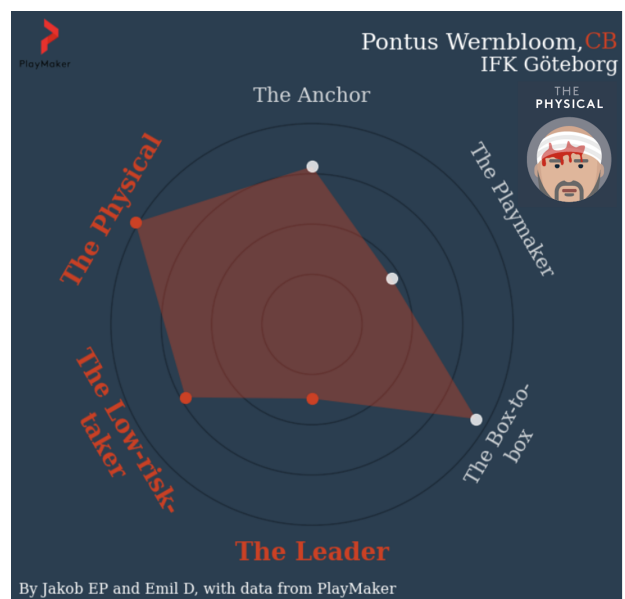
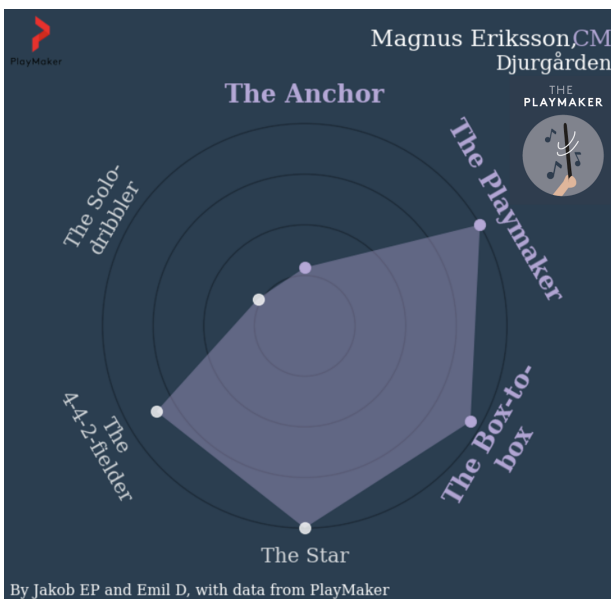


Figure 10: Left plot displays Magnus Eriksson, classified as "The Playmaker" in position 'CM', and how much he correlates to the playing styles in position 'OW'. Right plot displays Pontus Wernbloom, classified as "The Physical" in position 'CB', and how much he correlates to the playing styles in position 'CM'.

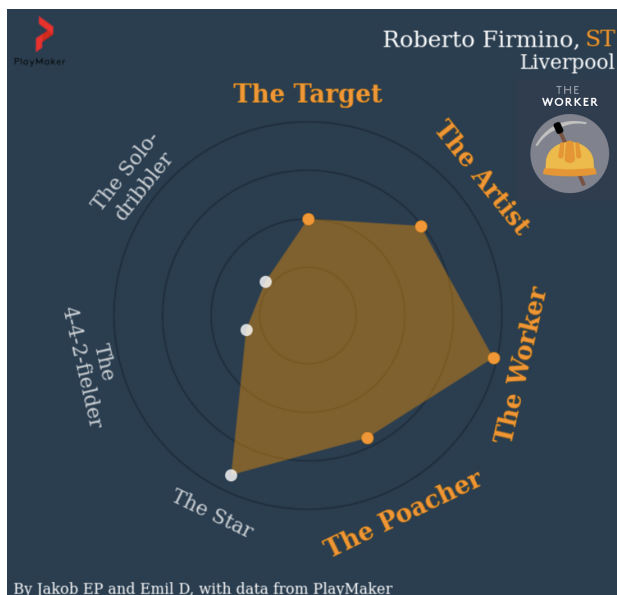


Figure 11: Left plot displays Roberto Firmino Eriksson, classified as "The Worker" in position 'ST', and how much he correlates to the playing styles in position 'OW'. Right plot displays Timo Werner, classified as "The Poacher" in position 'ST', and how much he correlates to the playing styles in position 'OW'.

5 Discussion

The three developed models from this master thesis show varying results with different interesting findings. The results of the position detection model show a relatively high accuracy, beating the available models at PlayMaker for both Allsvenskan and Premier League. Model v.1 show bad result in terms of accuracy against the validation data set, but still produced valuable findings when it comes to complexity of today's football players, something further discussed in Section 5.2. Model v.2 show good accuracy against the revised validation data set, but varies a lot between positions, with 'ST' having the best accuracy and 'CM' the worst. Further reasoning about this, together with other discussion points about Model v.2, can be found in Section 5.3.

5.1 Position detection model

The developed position detection model performed well on the validation data set with an accuracy of 0.84 for Allsvenskan and 0.89 for Premier league. One can argue that we already with the position detection model have achieved the main aim of characterising and classifying a footballer's playing style in a fully data-driven and objective manner. This comes from what is mentioned in Definition 1, the playing style of a player is much dependent on the areas of the pitch where the player performs the actions. Knowing whether a player is a Striker or a Full back is in some way enough of a playing style categorisation, depending on who you ask. However, since we, with our existing hypothesis in mind, wanted to go deeper the model development continued.

5.1.1 Player positions from formation data

Even though the results were good, position detection models probably belong to the past. Usually the positions of the players are a part of the match event data, in the form of formation data. This gives almost 100% accuracy and is more detailed and adjustable.

There are some merits of using a data-driven position detection model instead of using the formation data. By detecting the positions of the players in a data-driven way, we take into consideration where they actually are making their actions on the pitch, something that can differ from the registered lineups at times.

A data-driven position detection model also have the benefits of being more objective than labelling positions from lineups when it comes to certain formations in football. An example of this can be when it comes to the 4-3-3-formation. Some professionals of football might say that the two wide players playing up front are playing as wingers, while some might say they are playing as strikers. A post match interview with the two Liverpool players Mohammed Salah and Thiago Alcantara from April 2022 [20] further shows this type of discussion. Here, Thiago argues that Salah plays as a striker while Salah disagree and believes he is playing as a winger. Using the position detection model this discussion can be ended with the conclusion that Thiago is of the "correct" opinion.

5.2 Model v.1

The Model v.1 results, seen in Table 12, displays a model which only beats the baseline model accuracies for the playing position 'FB'. It gives an indication of how today's football players are too complex and versatile to be able to just be divided into either "Offensive" or "Defensive", with the fullbacks as a possible exception. This statement is also based on the difficulties with mapping our original hypothesis into "Offensive"/"Defensive"-classifications, where especially the playing styles "The Box-to-box" and "The Playmaker", for 'CM', can be argued to be considered both "Offensive" and "Defensive". However, alternating the mapping of those playing styles gave no better model results. So other than that Model v.1 could have some practical use when it comes to classifying offensive and defensive fullbacks it is quite limited. Due to these results of Model v.1 we can state that the hypothesis, saying multiple playing styles exists within each position, still holds. Another model had to be considered.

5.3 Model v.2

Evaluating our model is challenging as no objective ground truth exists for characterizing playing style. The revised validation data set must be considered to be the truth for any meaningful analyses of the playing styles certainties. This is of course highly subjective and as stated in Section 1.4 a delimitation the master thesis must take. The reason for that being how the field of football analytics is rather new, and with that suggestions on

other validation alternatives lacking (considering our niche approach). We can at least find comfort in that an unsupervised PCA-modelling approach makes the computations carried out (to find the playing styles) data driven, in an objective manner.

However, with Hypothesis 2, Assumption 5 and the above discussion in mind, the overall results of Model v.2 can be considered as satisfying in terms of accuracy and found playing styles. Also, the developed model managed, to a large extent, fulfills the aim and goals presented in Section 1.2.

5.3.1 Why this type of model?

The choice of model was decided by a trial and error process where several methods were taken into consideration. Using a clustering model, manually setting KPI weight patterns, or labelling more data to then use a supervised modelling approach were some of the methods/models tested.

What made the PCA-based model the go-to approach was how the findings from it well reflected the aims and goals, and especially the set up hypothesis for some of the positions. Furthermore, it also translates well to how the current structure of the PlayMaker.AI platform is built. By only using event data and KPIs, with Spiders such as Figures 10 and 11 as the model outputs, the model can be easily integrated to the current platform.

5.3.2 Certainty of found playing styles

An important thing to analyse about the developed model is how well the found playing styles translates to actual playing styles seen on the pitch. This is something hard to measure and get a clear grip of, but what can help in analysing this translations are the resulting accuracies in combination with the class metrics for each playing style.

For the position 'ST' the total accuracy is 0.79, which is the highest among all the positions from looking at Table 18. Within the Striker position the playing style "The Poacher" have the highest F1-score of 0.81. "The Worker" has the lowest F1-score, with a score of 0.76 according to Table 19. Both of the accuracies and the two F1-scores are relatively high and give us some indication of that the found playing styles exists both in the data, as well as on the pitch.

If we instead look at the position 'CM' with an accuracy of 0.64, worst of all the positions, we have three found playing styles which all are well known by football enthusiasts; "The Box-to-box", "The Playmaker" and "The Anchor". However, this is not something that the model recognises as well. All of the playing styles have F1-scores lower than 0.70 with "The Box-to-box" scoring as low as 0.56, see Appendix 3 for further results. This means that the found playing styles, even though they can be clearly differentiated by a football enthusiast, can not be as clearly differentiated by the developed model.

From the above reasoning, the certainty for each position and playing style, can be examined by combining the results of the accuracies with the class metrics. Hence, it can be said that the best modelled and certain playing style is "The Poacher", and the worst "The Box-to-box". This have to do with the developed model having it easier to differentiate between playing styles which have some on the ball actions that can be considered as typical for that playing style. For example, "The Poacher" takes many shots, scores many goals and does a lot of touches in the box. "The Target" win many challenges and makes many headers. However, for the playing style "The Box-to-box" it is much harder to exactly tell what those type of actions are, something also stated in the hypothesis in Section 1.5.

Some of the found playing styles are also highly correlated to players simply being good in many KPIs (see notebook [16] for weight patterns). Those playing styles are "The Leader", "The Star" and "The Artist". Also, good performing players often play in good performing teams and vice versa. Thus the model is not as good at finding for example "The Leader" in a less good team, something that is exemplified by Figure 10. Pontus Wernbloom is a player who many would consider to be a leader on the pitch. But because of the bad performance of his team IFK Göteborg last season, in combination with his own not so great season, the model results show that he is one of the centre backs in Allsvenskan that correlates the least with "The Leader".

As stated by the delimitations in Section 1.4 the model does not take into consideration how a player acts in situations without the ball. This is something that influence some playing styles more than others. Using only event data, it is hard to model playing styles doing much of their contribution to the team without the ball, such as "The Worker" and "The Box-to-box". A player that is known for hard work and making space for his teammates is Timo Werner. He is a great example of a player that scores a lot higher for the playing style "The Worker" if it was up to a football enthusiast to decide, instead of the developed model. But since making space-opening-runs for his teammates is not something included as a variable in the model, he gets assigned

"The Poacher".

For future users of the model results, knowing that there is differences in certainty among the playing styles is of great importance, and can also help making better informed decisions while using the results as part of any decision basis. As an example, if scouting for a central midfielder, the scout can not be as certain of the findings in comparison to if he or she scouts for any playing style within the striker position.

5.3.3 Usage of the model

The model can be used in several ways. Perhaps the most valuable usage of the model is how it manages to reduce the many KPIs to just a few playing styles that any ordinary football enthusiast can understand and recognise. This makes it easier for users of the PlayMaker.AI platform, that are not as interested in advanced football modelling, to search and scout for players in a simple way. From just looking at the playing style Spiders, the user gets a good picture of the players (see Figures 10 and 11 for examples).

Another usage of the model is that it can help with looking at how players would perform in different positions. In a scouting process this can be of great value as it can broaden the choices of players to look for in a position and desirable playing styles.

Finally, something that can not be underestimated is the improvement of the user experience for the PlayMaker.AI platform. Only looking at numbers, Spiders and heat maps can give an overwhelming amount of information. To reduce all the statistics into a colorful Avatar and/or an icon (as seen in the top right corner of Figures 10 and 11) contributes to a greater user experience and feel.

5.4 Future work

5.4.1 Playing styles of teams

Further work could be to investigate the playing styles of teams. From using a similar approach as in this master thesis, one could look at team-KPIs to see what different playing styles that exists in the data. A similar PCA approach as in Model v.2 or a clustering algorithm could be something worth trying out.

By combining information from playing style of a team, and the playing style of a player, a more complete player-profile can be created. This profile could be of great use in a platform such as PlayMaker.AI. Also, it would be interesting to investigate if there exists any patterns with team playing styles and what the squad consists of in terms of player playing styles. Is there any correlation between the set of playing styles in a team and the success of the team?

5.4.2 Manually setting KPI-weights

A somewhat similar approach to Model v.2 is to simply set the weight patterns manually (see Figure 8 for weight pattern example). By using the hypothesis in Section 1.5 as reference, the different weights of the KPIs can be set to arbitrary chosen values, preferably by someone with great domain knowledge. This approach ensures that we get the playing styles that we want from the hypothesis.

There is however a big drawback of using this approach. No matter how good of domain knowledge about football the author of the hypothesis, and the setter of the weight patterns have, there is always a possibility that the playing styles to look for are not reflected by the data. There are multiple factors that might affect this, for example; how the KPIs are computed, normalization techniques, the amount of data, what kind of data and the "quality" of the data.

An example of the above explained drawback can be seen when comparing the hypothesis in Section 1.5 with the found playing styles from Model v.2 in Section 4.3.2. One made change is the removal of the playing style "The #10" from the found playing styles. The reason for this is that the available data could not clearly separate between the more deep-lying playing style "The Playmaker", and the more advanced "The #10". This probably has to do with that they correlates well to each other in terms of important KPIs, see the hypothesis in Section 1.5. If the approach of manually setting the KPIs is instead used, the playing style "The #10" would exist, but highly dependent on the choices of the weights patterns and might not reflect the data.

5.4.3 Supervised modelling approach

From labelling the playing style of football players, a supervised approach with the labelled data as the dependent variable and KPIs as the independent variables, can also be used. Using an approach such as this requires a large data set and great domain knowledge of football from the ones labelling the data, preferably football experts, coaches and scouts.

A supervised approach would differ to the unsupervised model developed in this master thesis (Model v.2) in the sense that it would be "steered" towards the labelled data. An unsupervised approach, together with domain knowledge and analysing of the results, is better at telling us what the available data actually shows. A great example of this is the findings of the playing style 'The Target' from Model v.2. With a supervised model, with labelling as the hypothesis in Section 1.5, this playing style would not exist, while our developed model clearly shows that the playing style exists in the data.

6 Conclusions

The main aim with this master thesis was to characterise and classify a footballer's playing style in a fully data-driven and objective manner, with some more specific goals also set up. Overall, we can consider the specific goals as totally fulfilled and the main aim as partly achieved. With the goals met we can conclude; our resulting model fills a part of the problematic gap of translating football data statistics into common football knowledge, which the whole area of football analytics struggles with.

Following the above discussion and reasoning we can conclude that the final developed model (i.e. Model v.2) will be of great benefit included at the PlayMaker.AI platform, in the presented form displayed in Figures 10 and 11, when the descriptions of the playing styles in Tables 13 — 17 are added. In summary, an unsupervised PCA based model is suitable when it comes to identifying and classifying playing styles with the use of, on the ball, match event data.

Finally, coming back to the main aim being only partly achieved, that conclusion is based on how we throughout the master thesis utilized and worked against the validation data set, which by construction was created in a subjective manner. Hence, we can not state that the aim, which explicitly says to identify playing styles objectively, is fully met. However, as already mentioned in Section 5.3, we lacked alternatives and from that discussion we can at least conclude that the playing styles were found objectively, even though they were validated subjectively.

The important conclusion drawn in the paragraph above firstly leads us into wanting to investigate how to make football analytics less subjective and more objective. How could we "science(fy)" this area even more and depend less on evaluating models (as we have done) against subjective views? But, before even starting to discuss how this could be done we stopped and asked ourselves: Is this something we even want to go towards? So secondly it leads us into the discussion regarding what place football analytics should take in the football industry. Should it remain a valuable tool to complement the former players and journalists which we keep on calling the "experts"? Should it take over? Should perhaps football stay subjective, and we keep on trusting the experts over our own eyes and what we believe to be the problem as of why our favourite team keep on under-performing? Even though these are questions asked openly, without an obvious associated answer or conclusion attached to them, we decide to end with the concluding remark, and the most important conclusion of this master thesis: It is the subjectiveness of football which is what makes it so interesting, addictive and beautiful.

Bibliography

- [1] Van Haaren, J. State of the football analytics industry in 2021 [article]. SciSports. [updated 2021-03-17; read 2022-04-11] Available at: <https://www.scisports.com/state-of-the-football-analytics-industry-in-2021/>
- [2] Kidd, R. The World's Biggest Soccer Clubs Find A 'Smarter' Way To Scout Transfer Targets [article]. Forbes. [updated 2019-04-14; read 2021-09-24].
- [3] PlayMaker.AI [webbpage]. Football analytics as a service. [updated 2022; read 2022-03-16] Available at: <https://www.playmaker.ai/>
- [4] Jupyter [webbpage]. Jupyter Notebook. [updated 2022; read 2022-04-24] Available at: <https://jupyter.org/>
- [5] gitHub [webbpage]. avatar-playing-style. [updated 2022] Available at: <https://github.com/Sommarro-Devs/avatar-playing-style>
- [6] Lindholm, A., Wahlström, N., Lindsten, F., & B. Schön, T. *Supervised Machine Learning*. Draft version: April 30, 2021
- [7] Min-max scaler [webbpage]. Supervised learning. Scikit-learn 1.0.2. [updated 2022; read 2022-04-27]. Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>
- [8] Cattell, R. B. *The Scree Test For The Number Of Factors, Multivariate Behavioral Research*, 1(2):245–276. Volume 1, Issue 2: 1966
- [9] Possession-adjusted [webbpage]. Wyscout Glossary. [updated 2020; read 2022-03-16] Available at: https://dataglossary.wyscout.com/p_adj/
- [10] Rydén, J. *Stokastik för ingenjörer*. 2th ed. Lund: Studentlitteratur; 2015.
- [11] Lee, A. Choosing a Baseline Accuracy for a Classification Model [article]. Towards Data Science. [updated 2021-05-21; read 2022-04-11] Available at: <https://towardsdatascience.com/calculating-a-baseline-accuracy-for-a-classification-model-a4b342ceb88f>
- [12] Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D. & Giannotti, F. (2019). *A public data set of spatio-temporal match events in soccer competitions*. Scientific Data. 6. 10.1038/s41597-019-0247-7.
- [13] gitHub [webbpage]. avatar-playing-style, main_lib_position_detection.ipynb. [updated 2022] Available at: https://github.com/Sommarro-Devs/avatar-playing-style/blob/main/lib/position_detection.ipynb
- [14] gitHub [webbpage]. avatar-playing-style, helpers_lib.py. [updated 2022] Available at: https://github.com/Sommarro-Devs/avatar-playing-style/blob/main/lib/modules/helpers_lib.py
- [15] gitHub [webbpage]. avatar-playing-style, main_lib_model_v1.ipynb. [updated 2022] Available at: https://github.com/Sommarro-Devs/avatar-playing-style/blob/main/lib/model_v1.ipynb
- [16] gitHub [webbpage]. avatar-playing-style, main_lib_model_v2.ipynb. [updated 2022] Available at: https://github.com/Sommarro-Devs/avatar-playing-style/blob/main/lib/model_v2.ipynb
- [17] gitHub [webbpage]. avatar-playing-style, viz_lib.py. [updated 2022] Available at: https://github.com/Sommarro-Devs/avatar-playing-style/blob/main/lib/modules/viz_lib.py

- [18] gitHub [webbpage]. avatar-playing-style, models_lib.py. [updated 2022] Available at: https://github.com/Sommarro-Devs/avatar-playing-style/blob/main/lib/modules/models_lib.py
- [19] gitHub [webbpage]. avatar-playing-style, config.py. [updated 2022] Available at: <https://github.com/Sommarro-Devs/avatar-playing-style/blob/main/lib/modules/config.py>
- [20] Twitter [webbpage]. Thiago Alcantara and Mohammed Salah post match interview. [published 2022-04.19] Available at: <https://twitter.com/footballdaily/status/1516525000735535104>

Appendix A

1 KPI_org

Original Key Performance Indicators **KPI_org**, with a brief explanation, existing on PlayMaker.AI for each available player and used in this master thesis are the following:

- ◇ **goals** – Goals scored.
- ◇ **goals 90** – Goals scored per 90 minutes played.
- ◇ **ass** – Assists made.
- ◇ **ass 90** – Assists made per 90 minutes played.
- ◇ **points 90** – Goals and assists per 90 minutes played.
- ◇ **xg 90** – Expected goals per 90 minutes played.
- ◇ **sp xg 90** – Expected goals from set pieces per 90 minutes played.
- ◇ **xg impact** – Impact on teams xG-ratio (when on the pitch).
- ◇ **mean xg** – Average xG of shot taken.
- ◇ **shots 90** – Shots taken per 90 minutes played.
- ◇ **xa 90** – Expected assists per 90 minutes played.
- ◇ **dribbles 90** – Successful dribbles made per 90 minutes played.
- ◇ **dribbles %** – Success rate of dribbles made.
- ◇ **tib 90** – Touches in the box per 90 minutes played.
- ◇ **dze 90** – Danger zone entries per 90 minutes played.
- ◇ **tackles 90** – Successful tackles made per 90 minutes played.
- ◇ **tackles %** – Success rate of tackles made.
- ◇ **chall 90** – Successful challenges made per 90 minutes played.
- ◇ **chall %** – Success rate of challenges made.
- ◇ **g mist 90** – Mistakes made leading to goals per 90 minutes played.
- ◇ **mist 90** – Mistakes made per 90 minutes played.
- ◇ **int 90** – Interceptions made per 90 minutes played.
- ◇ **lost b 90** – Lost balls picked up per 90 minutes played.
- ◇ **dribb past 90** – Times dribbled past per 90 minutes played.
- ◇ **wb oh 90** – Won balls on opponent pitch half per 90 minutes played.
- ◇ **passes 90** – Successful passes made per 90 minutes played.
- ◇ **passes %** – Success rate of passes made.
- ◇ **xp %** – Success rate of expected pass, how likely a pass is to be successful.
- ◇ **crosses 90** – Successful crosses made per 90 minutes played.
- ◇ **crosses %** – Success rate of crosses made.
- ◇ **kp 90** – Successful key passes made per 90 minutes played.
- ◇ **pib 90** – Successful passes into the box made per 90 minutes played.
- ◇ **lb 90** – Successful long balls made per 90 minutes played.
- ◇ **lb %** – Success rate of long balls made.

- ◇ **directness %** – Rate of passes made directed towards the opponents goal.
- ◇ **avg pass dist** – Average distance of passes made.
- ◇ **avg keypass dist** – Average distance of key passes made.
- ◇ **gain 90** – Total distance the ball is driven towards the opponents goal per 90 minutes played.
- ◇ **headers 90** – Successful headers made per 90 minutes played.
- ◇ **headers %** – Success rate of headers made.
- ◇ **avg patch area** – The size of the average area covered defensively.

2 KPI_new

Newly computed Key Performance Indicators **KPI_new**, with a brief explanation, computed for each available player and used in this master thesis are the following:

- ◇ **xg share** – Contribution to/share of teams total expected goals.
- ◇ **shots share** – Contribution to/share of teams total shots.
- ◇ **xa share** – Contribution to/share of teams total expected assists.
- ◇ **dribbles share** – Contribution to/share of teams total dribbles.
- ◇ **tib share** – Contribution to/share of teams total touches in the box.
- ◇ **dze share** – Contribution to/share of teams total danger zone entries.
- ◇ **prog carries 90** – Progressive carries made per 90 minutes played.
- ◇ **prog carries share** – Contribution to/share of teams total progressive carries.
- ◇ **tackles 90 Padj** – Possession adjusted tackles KPI.
- ◇ **tackles share** – Contribution to/share of teams total tackles.
- ◇ **chall 90 Padj** – Possession adjusted challenges KPI.
- ◇ **chall share** – Contribution to/share of teams total challenges.
- ◇ **int 90 Padj** – Possession adjusted interceptions KPI.
- ◇ **int share** – Contribution to/share of teams total interceptions.
- ◇ **lost b Padj** – Possession adjusted lost balls KPI.
- ◇ **lost b share** – Contribution to/share of teams total lost balls.
- ◇ **dribb past 90 Padj** – Possession adjusted dribbled past KPI.
- ◇ **dribb past share** – Contribution to/share of teams total dribbled past.
- ◇ **wb oh 90 Padj** – Possession adjusted won balls opponent half KPI.
- ◇ **wb oh share** – Contribution to/share of teams total won balls opponent half.
- ◇ **fouls 90** – Fouls taken per 90 minutes played.
- ◇ **fouls 90 Padj** – Possession adjusted fouls KPI.
- ◇ **fouls share** – Contribution to/share of teams total fouls.
- ◇ **passes share** – Contribution to/share of teams total passes.
- ◇ **crosses share** – Contribution to/share of teams total crosses.
- ◇ **kp share** – Contribution to/share of teams total key passes.
- ◇ **pib share** – Contribution to/share of teams total passes into the box.
- ◇ **lb share** – Contribution to/share of teams total long balls.
- ◇ **directness share** – Contribution to/share of teams total directness.
- ◇ **gain share** – Contribution to/share of teams total gain.
- ◇ **headers share** – Contribution to/share of teams total headers.
- ◇ **avg patch area share** – Contribution to/share of teams total avg patch area.
- ◇ **off actions 90** – Number of offensive actions carried out per 90 minutes played. See Section 1 for which actions are considered to be offensive.

- ◇ **def actions 90** – Number of defensive actions (Possession adjusted) carried out per 90 minutes played. See Section 2 for which actions are considered to be defensive.
- ◇ **off actions ratio** – Ratio of offensive actions, i.e. $\text{off actions 90} / (\text{off actions 90} + \text{def actions 90})$.

All share KPIs above are computed according to Equation (6), described in Section 2.4. All Padj KPIs above are computed according to Equation (5), described in Section 2.3.

Appendix B

1 Offensive actions

Below the actions that count as offensive for Model v.1 are listed (named as in PlayMaker):

- ◇ **Passes accurate**
- ◇ **Passes (inaccurate)**
- ◇ **Dribbling**
- ◇ **Shots**
- ◇ **Left corners (accurate)**
- ◇ **Dribbles (Unsuccessful actions)**
- ◇ **Offsides**
- ◇ **Right corners (inaccurate)**
- ◇ **Assists**
- ◇ **Bad ball control**
- ◇ **Goals**
- ◇ **Dribbles (Successful actions)**
- ◇ **Set pieces crosses (inaccurate)**
- ◇ **Right corners (accurate)**
- ◇ **Left corners (inaccurate)**
- ◇ **Penalty**
- ◇ **Set pieces crosses (accurate)**
- ◇ **Direct free kicks (inaccurate)**
- ◇ **Lost balls**

2 Defensive actions

Below the actions that count as defensive for Model v.1 are listed (named as in PlayMaker):

- ◇ **Challenges (lost)**
- ◇ **Challenges (won)**
- ◇ **Interceptions**
- ◇ **Picking up free balls**
- ◇ **Tackles (Successful actions)**
- ◇ **Fouls**
- ◇ **Tackles (Unsuccessful actions)**

Appendix C

1 Results - Position detection model

1.1 Existing positional data/models

Table C.1: Confusion matrix PlayMaker estimated positions for Premier League 2021/2022.

	CB	FB	CM	OW	ST	nr of actual
CB	62	4	10	5	0	81
FB	0	20	1	39	0	60
CM	9	2	57	25	0	93
OW	1	4	1	53	3	62
ST	0	0	19	24	10	53
nr of predicted	72	30	88	146	13	349
Total accuracy: 0.58						

Table C.2: Class metrics PlayMaker estimated positions for Premier League 2021/2022.

	precision	recall	specificity	F1-score
CB	0.77	0.86	0.93	0.81
FB	0.33	0.67	0.87	0.44
CM	0.61	0.65	0.86	0.63
OW	0.85	0.36	0.96	0.51
ST	0.19	0.77	0.87	0.30

Table C.3: Confusion matrix PlayMaker primary positions for Premier League 2021/2022.

	CB	FB	CM	OW	ST	nr of actual
CB	71	2	6	3	0	82
FB	1	40	1	18	0	60
CM	5	1	76	11	0	93
OW	1	3	4	51	3	62
ST	0	0	15	16	22	53
nr of predicted	78	46	102	99	25	350
Total accuracy: 0.74						

Table C.4: Class metrics PlayMaker primary positions for Premier League 2021/2022.

	precision	recall	specificity	F1-score
CB	0.87	0.91	0.96	0.89
FB	0.67	0.87	0.93	0.75
CM	0.82	0.75	0.93	0.78
OW	0.82	0.52	0.96	0.63
ST	0.42	0.88	0.90	0.56

Table C.5: Confusion matrix PlayMaker estimated positions for Allsvenskan 2021.

	CB	FB	CM	OW	ST	nr of actual
CB	38	6	9	2	0	55
FB	1	8	1	31	0	41
CM	6	1	51	18	0	76
OW	0	1	9	46	3	59
ST	0	1	24	10	11	46
nr of predicted	45	17	94	107	14	277
Total accuracy: 0.56						

Table C.6: Class metrics PlayMaker estimated positions for Allsvenskan 2021.

	precision	recall	specificity	F1-score
CB	0.69	0.84	0.93	0.76
FB	0.20	0.47	0.87	0.28
CM	0.67	0.54	0.86	0.60
OW	0.78	0.43	0.92	0.55
ST	0.24	0.79	0.87	0.37

Table C.7: Confusion matrix PlayMaker primary positions for Allsvenskan 2021.

	CB	FB	CM	OW	ST	nr of actual
CB	43	5	8	1	0	57
FB	1	20	1	19	0	41
CM	4	1	57	12	2	76
OW	0	0	10	45	4	59
ST	0	1	19	8	18	46
nr of predicted	48	27	95	85	24	279
Total accuracy: 0.66						

Table C.8: Class metrics PlayMaker primary positions for Allsvenskan 2021.

	precision	recall	specificity	F1-score
CB	0.75	0.90	0.94	0.82
FB	0.49	0.74	0.92	0.59
CM	0.75	0.60	0.90	0.67
OW	0.76	0.53	0.93	0.62
ST	0.39	0.75	0.89	0.51

1.2 New Model

Table C.9: Confusion matrix position detection model for Premier League 2021/2022 season.

	CB	FB	CM	OW	ST	nr of actual
CB	71	6	0	0	0	77
FB	0	55	3	0	0	58
CM	2	1	74	15	1	93
OW	0	2	1	55	2	60
ST	0	0	0	4	49	53
nr of predicted	73	64	78	74	52	341
Total accuracy: 0.89						

Table C.10: Class metrics position detection model for Premier League 2021/2022 season.

	precision	recall	specificity	F1-score
CB	0.92	0.97	0.98	0.95
FB	0.95	0.86	0.99	0.90
CM	0.80	0.95	0.93	0.87
OW	0.92	0.74	0.98	0.82
ST	0.92	0.94	0.99	0.93

Table C.11: Confusion matrix position detection model for Allsvenskan 2021 season.

	CB	FB	CM	OW	ST	nr of actual
CB	57	2	0	0	0	59
FB	2	36	2	1	0	41
CM	4	1	60	8	3	76
OW	0	6	9	40	4	59
ST	0	0	0	4	42	46
nr of predicted	63	45	71	53	49	281
Total accuracy: 0.84						

Table C.12: Class metrics position detection model for Allsvenskan 2021 season.

	precision	recall	specificity	F1-score
CB	0.97	0.90	0.99	0.93
FB	0.88	0.80	0.98	0.84
CM	0.79	0.85	0.92	0.82
OW	0.68	0.75	0.92	0.71
ST	0.91	0.86	0.98	0.88

2 Results - Model v.1

Table C.13: Confusion matrix Model v.1 for position 'ST'.

	Offensive	Defensive	nr of actual
Offensive	49	18	67
Defensive	10	8	18
nr of predicted	59	26	85
Total accuracy:	0.67		

Table C.14: Class metrics Model v.1 for position 'ST'.

	precision	recall	specificity	F1-score
Offensive	0.73	0.83	0.31	0.78
Defensive	0.44	0.31	0.83	0.36

Table C.15: Confusion matrix Model v.1 for position 'CM'.

	Offensive	Defensive	nr of actual
Offensive	64	34	98
Defensive	33	3	36
nr of predicted	97	37	134
Total accuracy:	0.50		

Table C.16: Class metrics Model v.1 for position 'CM'.

	precision	recall	specificity	F1-score
Offensive	0.65	0.66	0.08	0.66
Defensive	0.08	0.08	0.66	0.08

Table C.17: Confusion matrix Model v.1 for position 'OW'.

	Offensive	Defensive	nr of actual
Offensive	46	17	63
Defensive	14	11	25
nr of predicted	60	28	88
Total accuracy:	0.65		

Table C.18: Class metrics Model v.1 for position 'OW'.

	precision	recall	specificity	F1-score
Offensive	0.73	0.77	0.39	0.75
Defensive	0.44	0.39	0.77	0.42

Table C.19: Confusion matrix Model v.1 for position 'FB'.

	Offensive	Defensive	#actual
Offensive	40	14	54
Defensive	17	17	34
nr of predicted	57	31	88
Total accuracy:	0.65		

Table C.20: Class metrics Model v.1 for position 'FB'.

	precision	recall	specificity	F1-score
Offensive	0.74	0.70	0.55	0.72
Defensive	0.50	0.55	0.70	0.52

Table C.21: Confusion matrix Model v.1 for position 'CB'.

	Offensive	Defensive	nr of actual
Offensive	13	21	34
Defensive	25	69	94
nr of predicted	38	90	128
Total accuracy:	0.64		

Table C.22: Class metrics Model v.1 for position 'CB'.

	precision	recall	specificity	F1-score
Offensive	0.38	0.34	0.77	0.36
Defensive	0.73	0.77	0.34	0.75

3 Results - Model v.2

Table C.23: Confusion matrix Model v.2 for position 'CM'.

	2.1	2.2	2.3	nr of actual
2.1	21	5	11	37
2.2	6	24	6	36
2.3	11	10	41	62
nr of predicted	38	39	58	135
Total accuracy:	0.64			

Table C.24: Class metrics Model v.2 for position 'CM'.

	precision	recall	specificity	F1-score
2.1	0.57	0.55	0.84	0.56
2.2	0.67	0.62	0.88	0.64
2.3	0.66	0.71	0.73	0.68

Table C.25: Confusion matrix Model v.2 for position 'OW'.

	3.1	3.2	3.3	nr of actual
3.1	23	3	7	33
3.2	8	20	1	29
3.3	5	4	23	32
nr of predicted	36	27	31	94
Total accuracy: 0.70				

Table C.26: Class metrics Model v.2 for position 'OW'.

	precision	recall	specificity	F1-score
3.1	0.70	0.64	0.83	0.67
3.2	0.69	0.74	0.87	0.71
3.3	0.72	0.74	0.86	0.73

Table C.27: Confusion matrix Model v.2 for position 'FB'.

	4.1	4.2	4.3	nr of actual
4.1	27	5	7	39
4.2	3	27	7	37
4.3	0	2	13	15
nr of predicted	30	34	27	91
Total accuracy: 0.74				

Table C.28: Class metrics Model v.2 for position 'FB'.

	precision	recall	specificity	F1-score
4.1	0.69	0.90	0.80	0.78
4.2	0.73	0.79	0.82	0.76
4.3	0.87	0.48	0.97	0.62

Table C.29: Confusion matrix Model v.2 for position 'CB'.

	5.1	5.2	5.3	nr of actual
5.1	29	2	3	34
5.2	3	23	0	26
5.3	16	14	38	68
nr of predicted	48	39	41	128
Total accuracy: 0.70				

Table C.30: Class metrics Model v.2 for position 'CB'.

	precision	recall	specificity	F1-score
5.1	0.85	0.60	0.94	0.71
5.2	0.88	0.59	0.97	0.71
5.3	0.56	0.93	0.66	0.70