Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Statistics

Bachelor Thesis

Data Analysis of Throw-ins in the Czech Football League

David Rozlivek

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

BACHELOR THESIS ASSIGNMENT

David Rozlivek

Business Administration

Thesis title

Data Analysis of Throw-ins in Czech Football League

Objectives of thesis

The object of a bachelor thesis is to assess set-pieces indicators (such as xG within amount of time, progressiveness, success rate, possession, and other indicators) from this type of set-pieces in Czech League. In addition, it will be evaluated a unique throw-in's passing probability model from Czech League. It will be used data exploratory analysis for set-pieces indicators and a gradient-boosted ensemble of decision trees for set-pieces passing model.

Methodology

There will be assessed chosen indicators which characterised set-pieces in the Czech Football League.

It will be used a gradient-boosted ensemble of decision trees (GBM) to evaluate Expected Passing Probability.

The proposed extent of the thesis

30 – 40 pages

Keywords

Decision trees, data exploratory analysis, football, set-pieces

Recommended information sources

KULLOWATZ, M. An Updated Expected Passing Model. In: American Soccer Analysis [online]. Available from <https://www.americansocceranalysis.com/home/2018/4/19/an-updated-expected-passing-model>

LARSON, R. – FARBER, E. *Elementary statistics : picturing the world.* Boston: Pearson Prentice Hall, 2015. ISBN 9780321693624.

MCKINLEY, E. Game of Throw-ins. In: American Soccer Analysis [online]. Available from <https://www.americansocceranalysis.com/home/2018/11/27/game-of-throw-ins>

STONE J., SMITH A., BARRY A. The undervalued set piece: Analysis of soccer throw-ins during the English Premier League 2018–2019 season. Available from <https://journals.sagepub.com/doi/full/10.1177/1747954121991447>

TUFFÉRY, S. Data Mining and Statistics for Decision Making. UK, West Sussex: Wiley, 2011. ISBN 978-0-470-68829-8.

YOUNG, J. Forget Everything You Think You Know About Goal Kicks. In: American Soccer Analysis [online]. Available from <https://www.americansocceranalysis.com/home/2016/12/7/attention-goalkeepers-stop-kicking-those-long-balls>

Expected date of thesis defence

2021/22 SS – FEM

The Bachelor Thesis Supervisor

Ing. Tomáš Hlavsa, Ph.D.

Supervising department

Department of Statistics

Electronic approval: 10. 2. 2022

prof. Ing. Libuše Svatošová, CSc.

Head of department

Electronic approval: 21. 2. 2022

doc. Ing. Tomáš Šubrt, Ph.D.

Dean

Prague on 15. 03. 2022

Declaration

I declare that I have worked on my bachelor thesis titled "Data Analysis of Throw-ins in the Czech Football League" by myself and I have used only the sources mentioned at the end of the thesis. As the author of the bachelor thesis, I declare that the thesis does not break any copyrights.

In Prague on 14th of March ______________________________

Acknowledgement

I would like to thank Ing. Tomáš Hlavsa, Ph.D. for his professional help. I would also like to thank Dr. Stylianos Kampakis, who is my data science mentor. Next, I would like to mention my family who supported me through my data science development. Most of all, I would like to say a big thanks to Wyscout, a data provider company, that provided me with the data.

Data Analysis of Throw-ins in the Czech Football League

Abstract

The increased significance of throw-ins at top-league clubs has recently been seen in practice, as foreign clubs have begun to recruit trainers who are only focused on throw-ins. Thus, this thesis will analyse throw-ins in general, especially how each attribute (e.g., throw-in distance, timing of taking a throw-in, throw-in angle, etc.) influences the completeness of a throw-in, possession retention after a throw-in, and scoring probability following a throw-in. Throw-ins are analysed using the exploratory data analysis and the construction of a basic machine learning model. The examined dataset contains approximately 80,000 throw-ins taken in the Czech Football League from 2015 to 2021. Data analysis indicates that the faster throw-ins are more likely to be successful and possession held. The shorter throw-ins are more likely to be successful and possession retained. Backward throw-ins have bigger probability of being successful and possession kept in comparison to forward throw-ins. A throw-in with a switch, which is taken in defensive or middle zone, has a better likelihood of scoring a goal and also avoiding a conceded goal.

Keywords: Football, Data Analysis, Decision Tree, Set pieces, Throw-ins, Completion, Possession retention, Expected Goals, Machine Learning, SHapley Additive exPlanation, LightGBM model.

Datová Analýza – Vhazovaní v České fotbalové lize

Abstrakt

Zvýšený význam vhazování v top zahraničních klubech se v poslední době projevuje i v praxi, protože zahraniční kluby si začaly najímat trenéry, kteří se zaměřují pouze na vhazování. Tato práce se tedy bude zabývat obecnou analýzou vhazování v České fotbalové lize, zejména tím, jak jednotlivé atributy (např. vzdálenost, načasování provedení vhazování, progresivita vhazování atd.) ovlivňují úspěšnost vhazování, zisku míče po vhazování a pravděpodobnost skórování po vhazování. Zkoumaná data obsahují přibližně 80000 vhazování provedených v České fotbalové lize v letech 2015 až 2021. Analýza je provedena obecnou průzkumnou analýzou dat a sestavením jednoduchého statistického modelu. Z analýzy dat vyplývá, že rychlejší provedení vhazování zvyšuje pravděpodobnost, že vhazování úspěšně najde spoluhráče a jeho tým udrží míč po vhazování. Zatímco vhazování s delší vzdáleností hodu snižuje pravděpodobnost, že vhazování nalezne spoluhráče a udrží balón po určitou dobu po vhazování. Vhazování směrem k vlastní brance má větší pravděpodobnost úspěšnosti a udržení míče ve srovnání s vhazováními směrem k soupeřově brance. Vhazování s dlouhou horizontální přihrávkou (tzv. switch), které je provedeno v obranném nebo středním pásmu, zvyšuje pravděpodobnost na vstřelení gólu a také na zabránění inkasovanému gólu.

Klíčová slova: Fotbal, Průzkumná analýza dat, Rozhodovací stromy, Standardní situace, Vhazování, Kompletnost, Držení míče, Očekávané góly, Strojové učení, SHapleyho vysvětlení modelu, LightGBM model.

Table of content

1 Introduction **10**

2 Objectives and Methodology **11**

2.1 Objectives 11

2.2 Methodology 11

2.2.1 Examined Dataset 11

2.2.2 Limitation of Examined Dataset 14

2.2.3 Descriptive Statistics 14

2.2.4 Clustering method 14

2.3 Ensemble Machine learning technique – LightGBM 15

2.3.1 Model Description 15

2.3.2 Hyperparameters Description 16

2.3.3 Evaluation of the Model Performance 18

2.3.4 Explanation of the Model 19

3 Introduction to Football Analytics **21**

3.1 Throw-in Definition 21

3.2 Machine Learning Models in Football 21

3.3 Terms Glossary 23

4 Practical Part **26**

4.1 Exploratory Data Analysis (EDA) 26

4.1.1 Throw-in End Location 27

4.1.2 Throw-in Distance 30

4.1.3 Throw-in Angle 32

4.1.4 Timing of Taking Throw-ins 34

4.1.5 Throw-ins Combos 35

4.1.6 Throw-ins with a Switch 37

4.1.7 Attacking Throw-ins 40

4.1.8 Throw-ins terms of Game State 42

4.1.9 Throw-ins terms of Player Difference 45

4.1.10 Feature Analysis Summary 47

4.2 Expected Throw model 51

4.2.1 Feature Engineering and Model Selection 51

4.2.2 Fitting model 51

4.2.3 Model Explanation 54

5 Conclusion **57**

6 References **58**

7 List of figure, tables, and abbreviations 60

7.1 List of figures 60

7.2 List of tables 61

7.3 List of abbreviations 61

1 Introduction

Football is a business like any other industry, where they use data analytics and data science to work more efficiently. During the match, television channels provide us basic data such as a shot count, possession time, etc. These numbers represent only a portion of the information that can be obtained from this game. There are several other statistical indications, as are demonstrated in the following chapters. The data is gathered by third-party companies such as Wyscout (Wyscout Spa, 2022). These companies provide us with event data (every action with the ball on the pitch) or tracking data (every movement on the pitch). Without proper data management, nothing is gained, which is why football clubs and other organizations employ data analysts and data scientists. (Soccerment Research, 2021).

Liverpool FC is one of the best examples of using data analytics to be effective at both sides – on the pitch and the market. They appointed a data analyst, Michal Edwards, as head of performance and analysis. He created a four-man research team, which was revolutionary at that time. The key factor is the current coach Klopp, who works closely with a research team. They have increased the market value of many players and improved team performance. They also have won the Premier League and the Champions League in 2019 (Soccerment Research, 2021).

Unpredictable moments contribute to football's popularity as a spectator sport. Numerous difficult-to-predict variables must be taken into account by analysts. This does not stop them from developing many models and research analyses that will assist us in forecasting the outcome of this game in order to protect it from becoming too chaotic. Innumerable articles of analysing football set pieces have been published. During training sessions, coaches pay less attention to throw-ins than they do to other types of set pieces. According to Wyscout Data, each match has an average of 54 throw-ins in the Czech Football League. Additionally, the proportion of throw-ins out of all set piece types is nearly equal to 46%. This demonstrates the high occurrence and the potential importance of throw-ins in terms of team success.

2 Objectives and Methodology

2.1 Objectives

The bachelor thesis focuses on the data analysis of an undervalued set piece (throw-ins) it evaluates the completion, possession retention, and creation of scoring opportunities of throw-ins in the Czech Football League since 2015. The objective is to comprehend the significance of various features in classifying the outcome of the target variable, which is the throw-in completion. The second objective is to develop a basic machine learning model that will enable us to better understand how each variable influences the prediction of the throw-in completion.

2.2 Methodology

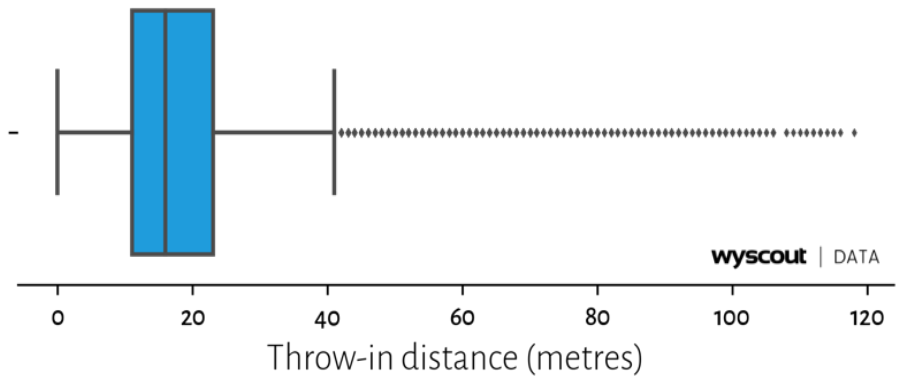
In this Section, the examined dataset and utilised approaches (in Section 4) will be presented and how they might be applied to accomplish the stated objectives. The following procedures are performed in a Python 3.7 environment.

2.2.1 Examined Dataset

The data, which is used for this bachelor thesis, was collected by Wyscout. Wyscout is a provider of performance data (Wyscout, 2022). They have been collecting data since the season of 2014/2015. Therefore, the dataset is a fairly large sample that contains data (features) of every event on the pitch over the last 6 seasons of the Czech Football League. There are 16 teams playing in this league, except for the season 2019/2020 when 18 teams played due to the pandemic situation. The raw Wyscout dataset contains around 150 features, including information on each action with the ball on the pitch, such as the location (x, y-coordinates), duration, the players' and teams' names, a few specific metrics (e.g., Expected Goals), and more comprehensive information about individual actions (the type of pass and duel, angle of the pass, distance, path of the attack), etc (Wyscout, 2022). Most of them are not valuable information for analysing throw-ins. On the other hand, a few relevant features are computed from the raw Wyscout dataset (such as progressiveness, switches, player and game state, home indicator, time since the previous action, etc.).

Data errors frequently occurred as a result of a lack of video proof, as evidenced by what we could observe on the Wyscout platform. The longest throw-in (male) distance is 59.817 meters, which was made on 21 April 2019 (Guinness World Records, 2019). Thus, exactly 4639 samples were excluded from 84,605 throw-in samples in our dataset due to being an unrealistic outlier with a relatively large pass distance (more than 40 metres), coordinates indicating strange end locations, or missing thrower's name. In Figure 1, a box plot shows the distribution of throw-ins distance via quartiles. Outliers is represented as small diamonds outside the box plot's whiskers computed by an inter-quartile function.

Figure 1: The Throw-in Distance Box Plot



Source: own elaboration; Data: Wyscout Data

To summarize, we train the following characteristics using a machine learning model on approximately 80000 throw-ins taken in the Czech Football League from 2015 to 2021, while excluding distance outliers (throw-ins greater than 40 metres), in order to classify a binary target variable referred to as outcome value with 0 (represents an unsuccessful throw-in) and 1 (represents a successful throw-in) labels:

Table 1: Trained features by the LightGBM model

Feature	Type	Units	Range
x	location	coordinate	(0,120)
y	location	coordinate	(0,80)
pass_endX	location	coordinate	(0,120)
pass_endY	location	coordinate	(0,80)
throw-in angle	geometry	degrees	(-180,180)
throw-in distance	geometry	metres	(0,40)
end distance from the opp's goal	geometry	metres	(0,125)
time from the previous action	duration	seconds	(0,485)
home indicator	context	category	[home, away]
game state	context	category	[losing, drawing, leading]
player difference	context	category	[-1, 0, 1]

Source: own elaboration

The start location is marked by the **x- and y-coordinates**, while the end location is denoted by **pass endX and pass endY**. Additionally, Wyscout data includes **pass angle** and **pass distance** for throw-ins. Despite this, **throw-in distance** can be calculated using Euclidean distance on those four coordinates for each throw-in. **End distance from the opposition's goal** determines how far the receiver of the throw-in is situated from the opposition goal, it is computed using Euclidean distance as well as throw-in distance. Arctangent function can be used to compute **throw-in angle** on those four coordinates. Arctangent computes angle between two vectors (dx, dy), dx is represented by subtracting pass_endX and x-coordinate, and dy is represented by subtracting pass endY and y-coordinate. **Time from the previous action** is simply computed by subtracting match time of throw and match time of previous action, which means how quick the player takes the throw-in**. Home indicator** specifies whether the thrower plays at home or away. **Scoring difference** implies the game state if the thrower is leading, drawing, or losing in the match. The feature "**Player difference**" refers to if the thrower's team has an extra man advantage over their opponent, is equal, or has one less player thanks to a red card.

2.2.2 Limitation of Examined Dataset

We do not expect the model to be perfect due to an insufficient number of relevant characteristics, such as throw-ins under pressure or other players' coordinates when we work with event data. Working with tracking data, where the velocity of the ball and players is additionally gathered compared to events data, would be even better for model construction. For instance, if the thrower increases the throw-in distance, which normally suggests a higher probability of taking an incomplete throw-in, but the teammate (receiver) is so far away from the opponent, thus, the likelihood of a throw-in being completed would increase. The distance between receiver of the throw-in and opponent is closely related with the pressure data, which would also help with enhancing the model. However, the thesis's goal is to understand throw-ins' characteristic patterns using Wyscout events data, which **Shapley Additive Explanations (SHAP)** plots via the LightGBM model enables us to investigate (as described in detail in Section 2.3.4).

2.2.3 Descriptive Statistics

The goal of descriptive statistics is to describe and organize a specific data set. A specific data set includes variables, which are divided into quantitative and qualitative data. Quantitative data has numerical variables (e.g., throw-in distance). Qualitative data represents labels, names, or number code (e.g., team name). Among other things, descriptive statistics components include the mean, median, skewness, correlation, and variance (Goos & Meintrup, 2015).

2.2.4 Clustering method

The OPTICS clustering approach is utilised, which is an algorithm for identifying density-based clusters in spatial datasets. Additionally, it is simple to utilise because it is not required to define a number of clusters. The OPTICS algorithm has two primary parameters: Eps and MinPts. While Eps specifies maximum radius of the neighbourhood, MinPts sets minimum number of points in an Eps-neighbourhood of that point. OPTICS operates similarly to an extended DBSCAN algorithm in principle, except that it supports an unlimited number of distance parameters Eps_i that are less than the generating distance Eps.

The main distinction is that it does not offer cluster memberships but instead records the order in which points are processed (the clustering order) and the following two pieces of information that would be utilised to assign cluster memberships by an extended DBSCAN algorithm (Aggarwal & Reddy, 2013). In the data analysis of throw-ins, we use the OPTICS algorithm to identify groups of throw-ins with three pass combinations.

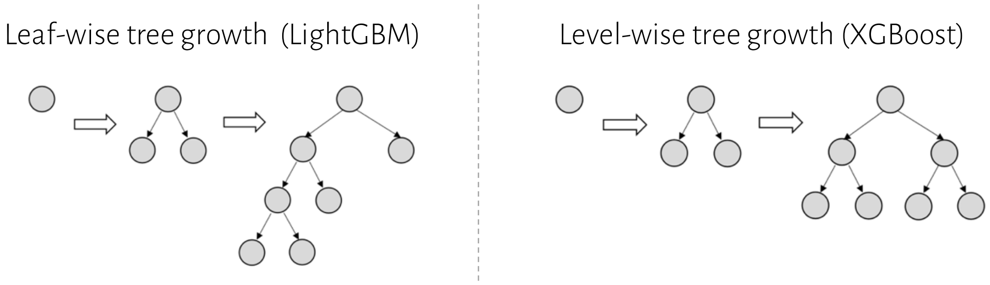
2.3 Ensemble Machine learning technique – LightGBM

2.3.1 Model Description

Microsoft (Microsoft Corporation, 2022) introduced an open source **the LightGBM model** (Ke, et al., 2017). It is a decision-tree-based technique that allocates the parameters in the input layer. The LightGBM model's primary characteristic is that it employs **leaf-wise tree growth** rather than the more often utilised level-wise tree growth to accelerate training (see Figure 2). In contrast to the level-wise tree growth method, the leaf-wise tree growth technique develops the tree around the node that results in the greatest error reduction. The leaf-wise tree growth algorithm may greatly decrease the number of tree nodes in contrast to the level-wise tree growth method, resulting in a considerable acceleration of the training process when the dataset is huge. Due to insignificant amount of data, leaf-wise may result in over-fitting, which is why LightGBM provides the max depth parameter to limit tree depth. Nonetheless, trees still develop leaf-wise even when max_depth is defined (Microsoft Corporation, 2022).

In this thesis, we use two LightGBM model training algorithms: **histogram-based (HB)** and **gradient-based one-side sampling (GOSS)**. The HB algorithm transforms the sorted dataset of parameters in the input layer into a histogram with a given number of data bins. As result, this approach uses much less memory while significantly increasing training speed. The GOSS approach randomly ignores data instances with modest gradients, while data instances with bigger gradients are more significant for computing information gain, implying that GOSS may produce pretty precise predictions while simultaneously lowering the quantity of the data (Ke, et al., 2017).

Figure 2: Types of Tree Growth


Leaf-wise tree growth (LightGBM)
Level-wise tree growth (XGBoost)

Source: own elaboration; inspired from (Microsoft Corporation, 2022)

2.3.2 Hyperparameters Description

A hyperparameter is a setting that is made before the start of the learning process. Hyperparameters are adjustable and may have a significant impact on a model training (DeepAI, 2022). In Table 2, used hyperparameters are listed along with their explanations, ranges, and default values. The following are two frequently used ways for optimizing hyperparameters. The typical approach of optimizing hyperparameters is **Grid Search (GS)**. This method finds the best performing hyperparameter from a collection of manually provided hyperparameters. Second method for optimizing hyperparameters is **Random Search (RS)**. This technique is almost identical to GS. However, RS uses random search instead of exhaustive search. The disadvantage of RS is that it cannot guarantee the optimal parameter combination because not all parameter values are tested (Bergstra & Bengio, 2012).

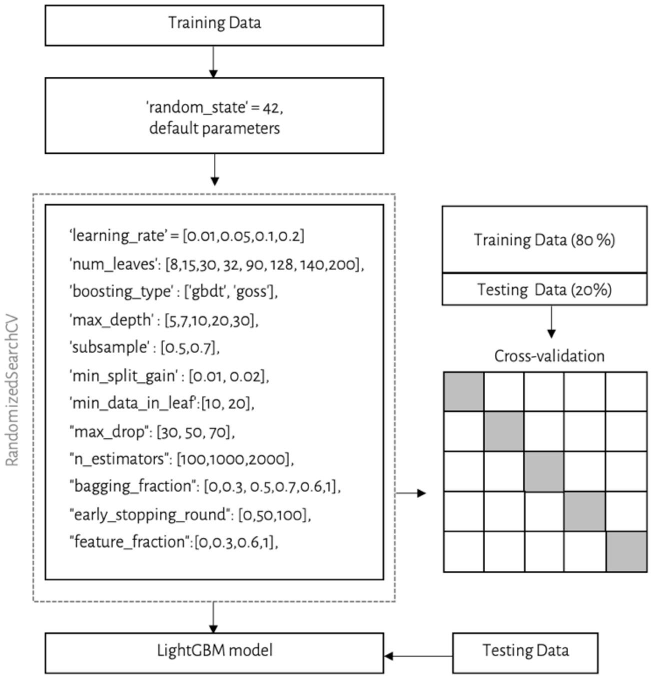
We have a relatively large data set and set of predefined hyperparameters. Therefore, we use RS as a tool to find the best potential set of hyperparameters for the model, which is less time-consuming than GS. We defined a f1-score as the scoring criterion function of the RS to measure the model performance. Furthermore, 5-fold cross-validation was used to improve the set of parameters. The flow diagram shows the whole process to find the best option of constructing the model with ideal option of set of hyperparameters (see Figure 3).

Table 2: The model's primary hyperparameters employed in this research

HYPERPARAMETERS	DESCRIPTION	TYPE	RANGE	DEFAULT
boosting_type	Boosting method	enum	[gbdt, rf, goss, dart]	gbdt
n_estimators	number of boosting iterations	int	>=0.0	100
learning_rate	The shrinkage rate	double	>0.0	0.1
num_leaves	Maximum leaves for each trained tree	int	>=0.0	31
max_drop	Maximum amount of data intervals	int	No limit	50
max_depth	Maximum depth of each trained tree.	int	No limit	-1
min_data_in_leaf	Prune by minimum number of observations requirement.	int	>=0.0	20
min_split_gain	the minimal gain to perform split	double	>=0.0	0
early_stopping_round	The maximum number of rounds without progress before we should call it quits	int	>=0.0	0
bagging_fraction	Percentage of rows used per iteration frequency	double	(0,1)	1
feature_fraction	Percentage of the selected parameters before training each tree	double	(0,1)	1

Source: own elaboration from (Microsoft Corporation, 2022)

Figure 3: Flow diagram of optimizing set of hyperparameters



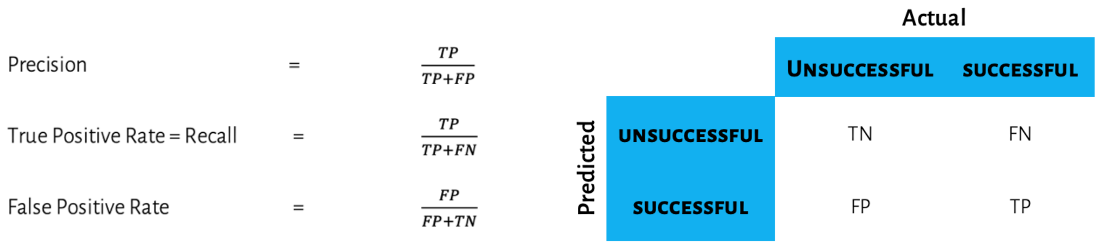
```
'learning_rate' = [0.01,0.05,0.1,0.2]
'num_leaves': [8,15,30, 32, 90, 128, 140,200],
'boosting_type' : ['gbdt', 'goss'],
'max_depth' : [5,7,10,20,30],
'subsample' : [0.5,0.7],
'min_split_gain' : [0.01, 0.02],
'min_data_in_leaf':[10, 20],
"max_drop": [30, 50, 70],
"n_estimators": [100,1000,2000],
"bagging_fraction": [0,0.3, 0.5,0.7,0.6,1],
"early_stopping_round": [0,50,100],
"feature_fraction":[0,0.3,0.6,1],
```

Source: own elaboration; inspired by: (Gan, et al., 2021)

2.3.3 Evaluation of the Model Performance

For evaluation of a binary decision problem, we utilise **Receiver Operating Characteristic (ROC)** and **Precision-Recall (PR)** curves, which originated from a confusion matrix. As shown in Figure 4, **the confusion matrix** consists of four elements: True positives (TP) are instances that are appropriately categorized as such. False positives (FP) are negative cases that have been wrongly classified as positive. True negatives (TN) are negatives that are accurately identified as such. Finally, false negatives (FN) are positive cases that have been mistakenly categorized as negative. **The False Positive Rate (FPR)** is plotted on the x-axis. This computes the proportion of negative examples that are incorrectly classified as positive. On the other hand, **the True Positive Rate (TPR)** is plotted on the y-axis in ROC space. TPR implies the proportion of correctly identified positive examples. Recall lies on the x-axis of PR curve and precision falls on the y-axis. While recall is same to TPR, precision indicates the proportion of samples identified as positive that are actually positive. Each metric is defined also in Figure 4 (Davis & Goadrich, 2006).

Figure 4: Frequently Used Machine Learning Evaluation Metrics



Precision $= \frac{TP}{TP+FP}$

True Positive Rate = Recall $= \frac{TP}{TP+FN}$

False Positive Rate $= \frac{FP}{FP+TN}$

		Actual	
		UNSUCCESSFUL	SUCCESSFUL
Predicted	UNSUCCESSFUL	TN	FN
	SUCCESSFUL	FP	TP

Source: own elaboration; inspired by: (Davis & Goadrich, 2006)

2.3.4 Explanation of the Model

SHapley Additive exPlanation (SHAP) is a tool that may aid with visualising of the explanation of the outputs of machine learning models (Lundberg, 2018).

SHAP values are derived from coalitional game theory that suggest how to allocate the pay-out among the features. This approach retrains the model on all feature subsets, giving a significance value to each feature indicating its impact on the model prediction. After training the model with the current feature, another model is trained without it. The predictions of these two models are compared $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$. After that, The SHAP values are calculated and applied as features attributions. Due to the fact that the impact of removing a feature is dependent on the presence of other features in the model, the previous differences are evaluated for all feasible subsets. The greater the absolute SHAP value, the more significant is the feature (Lundberg & Lee, 2017). To determine the global significance of a feature, we calculate an average of the absolute SHAP values for each feature throughout the data:

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S (x_S)],$$

where F symbolises all features, $S \subseteq F$ represents subsets of all features, $f_{S\cup\{i\}}$ is a model trained with the current feature, and f_S indicates the model trained without the current feature (Lundberg & Lee, 2017).

Additive feature attribution methods utilises an explanation model that is a linear function *f(x)* of binary variables based on a single input *x*. The explanation model often makes use of simplified inputs *x'* that are converted to the original inputs through the function *x = hx(x')*, because it is not simple to interpret the original model. This technique makes an attempt to guarantee that *g(x') ≈ f(hx(z'))*, when *z ≈ x'* (Lundberg & Lee, 2017). This technique is defined by the following formula:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i \, z'_{i,}$$

where $z^{'} \in \{0, 1\}^{M,}$ M means the number of simplified input features, and $\phi_{\mathrm{i}} \in$ R (Lundberg & Lee, 2017).

SHAP values are used to describe the global importance of characteristics, and also to interpret the significance of features on an individual throw-in. Ribeiro et. al. introduced the concept **"Local Interpretable Model-agnostic Explanations" (referred to as LIME)** for explanations how individual throw-ins change (Lundberg & Lee, 2017). The explanation model *g(z')* is created by minimizing a loss function *L* and penalizing the complexity of *g*, it is weighted by the local kernel π_X as follows:

$$\varepsilon = \arg\min L(f, g, \pi_{x\prime}) + \Omega(g) \,,$$

where the initial model is denoted by *f*, the kernel function is denoted by $\pi_{X'}$, and the regularization constant is denoted by *Ω* (Lundberg & Lee, 2017).

3 Introduction to Football Analytics

The next paragraphs discuss the most often utilised football machine learning models and their associated indicators in order to facilitate comprehension of the bachelor thesis results. All of these statistical indicators are used to characterize the team's performance in throw-ins. However, first and foremost, the definition of a throw-in is introduced.

3.1 Throw-in Definition

The term "**throw-in**" describes a football situation when the ball is returned to play by a thrower with their hands. Throw-ins occur when the referee stops the play because the ball leaves the pitch over the side-line (International Football Association Board, 2021). **A throw-in** is one out of five distinct types of set pieces in football. Other types are penalties, free-kicks, corners, and goal-kicks. At the moment of the throw-in, the player must have both feet on the ground behind the out-of-play side-line and the ball must be lifted over his head with both hands and then be delivered back into play. The player out of possession must stay at least two meters in front of the player who delivers the ball (International Football Association Board, 2021).

3.2 Machine Learning Models in Football

Evaluating the likelihood of a successful pass, the chance of a goal being scored, or any other predictive model in football is now a vital aspect of football analytics. Thus, there are already several prediction models that aid in football decision-making. This chapter describes a selection of models and their creators.

Expected Goals (xG) model, which was introduced by Sam Green, is a predictive machine learning model used to measure the quality of a chance by calculating the likelihood of scoring for every shot made in the game (Whitmore, 2021). When we calculate the xG of a shot, the xG model computes the probability of scoring based on several factors (parameters): the location of the shot; the location of the assist; shot type; assist type; presence of a dribble of a shooter before the shot; game statement (open play/set piece); transition statement (positional/counterattack); and the tagger's assessment of the danger of the shot. All these parameters (plus a few technical ones) are used to train the xG machine learning model on the historical data and predict the probability of the shot being scored. A greater probability indicates a greater likelihood of scoring in terms of the specified parameters. A shot with xG value of 0.3 is scored in 30% cases based on all the parameters. Penalties are easier to estimate. According to Wyscout data, the probability of scoring a penalty is about 76 %. Thus, the penalty xG value has been set to 0.76. **Expected Goals against (xGA)** is the probability that an opponent's shot results in a goal based on the attributes given above. For simplicity, an opponent's xG is assigned as xGA. **Expected Goals difference (xGD)** is calculated by subtracting the xG value from the xGA value (Wyscout, 2021). In the practical part of this thesis, we use the xG model to assess the throw-ins performance.

Matthias Kullowatz developed the **Expected Pass (xPass)** machine learning model. xPass calculates the probability of a pass being successful. A successful pass is a pass which directly finds the teammate of a passer. To be specific, this model assigns the likelihood to each pass based on several parameters from the dataset: a location of the passer, location of the receiver, angle of the pass, distance of the pass, type of the pass (longball/throughball), etc (Kullowatz, 2018).

The **Expected Throw (xThrow)** and **Expected Retain (xRetain)** models are used to predict the chance of a throw-in being complete and the probability of retaining possession following a throw-in based on specific criteria. Eliot McKinley designed these models, which are based on the xPass concept but include a few additional capabilities (parameters). The models used in this bachelor thesis will be almost identical to those used in Eliot McKinley’s article but will be trained on a different dataset from a different league, as the objective is to understand the set piece patterns in the Czech Football League. Additionally, we will create a benchmark for the model that excludes Eliot's article in order to determine how the features used to train the model are indeed relevant for predicting whether the throw-in will be completed (McKinley, 2018).

One of the components of this bachelor thesis is the trend analysis. Three writers, led by Joseph Antony Stone, previously published a similar piece of work in which they analysed throw-ins in the Premier League in terms of their success and ball possession percentage, and chance creation after throw-ins (Stone, et al., 2021).

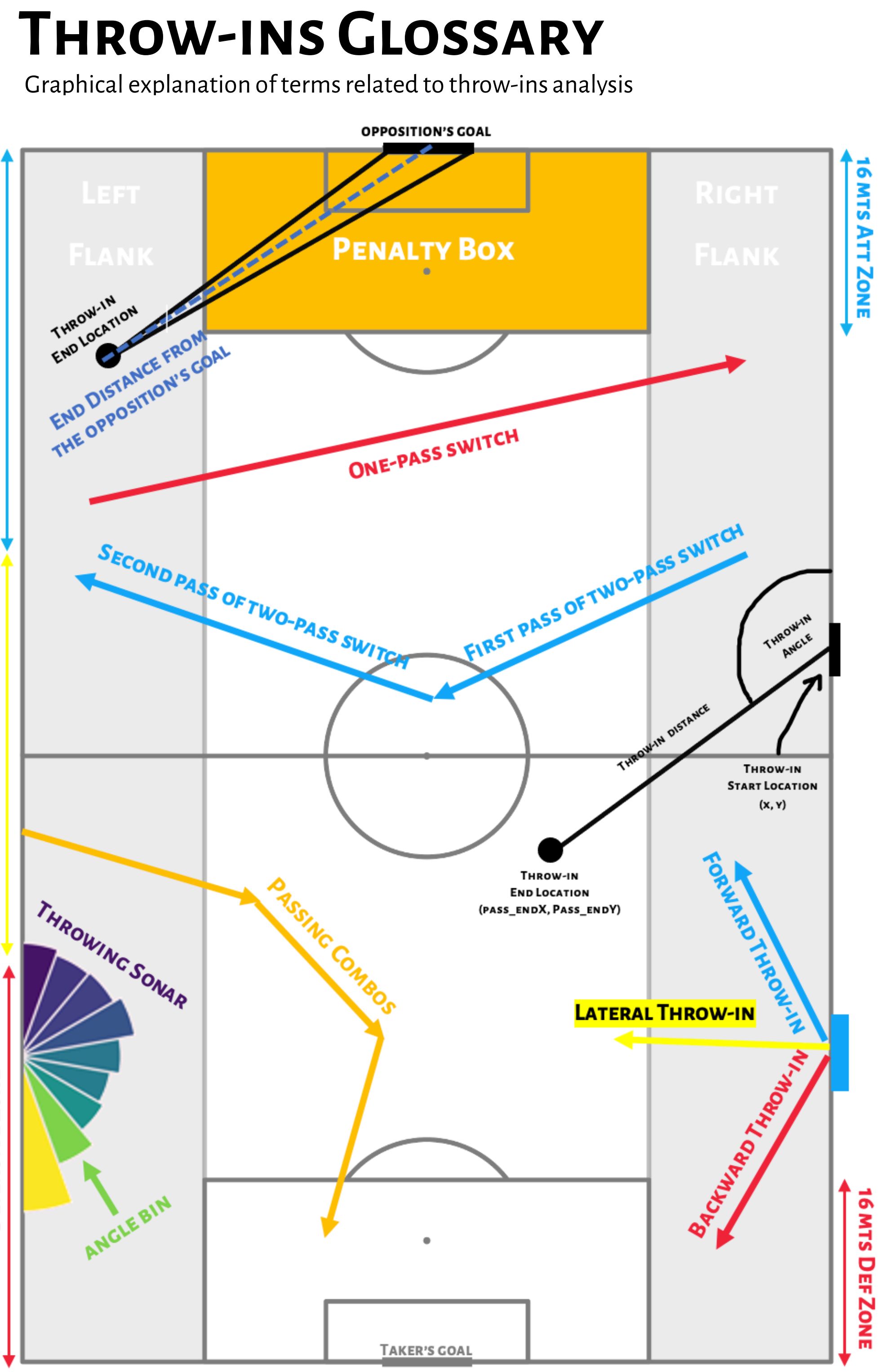
3.3 Terms Glossary

In this section, the meaning of statistical indicators, features and terms related to throw-ins, used in data analysis and subsequently in building predictive model, are described. In this thesis, the most used statistical indicators are completing of a throw-in, possession retaining after the throw-in, and creating chances from the throw-in (via xG). Time frames for the following statistical indicators are set based on domain knowledge:

- **A completed (successful) throw-in** is a throw-in that directly passes to a teammate without being touched by an opponent.
- **A possession retained throw-in** is defined as one that does not have to aim directly at a teammate, but then the team that has returned the ball into play possesses the ball again within seven seconds of the throw-in being taken.
- **A throw-in with a shot** is a throw-in in which the team taking the throw-in creates a scoring opportunity within 15 s of taking the throw-in. On the other side, **a throw-in with an opposition's (opponent's) shot** is one in which the team defending the throw-in creates a scoring opportunity within 15 s of taking the throw-in.
- **xG from throw-in** - as previously stated, each shot has an xG value; consequently, we can also evaluate the throw-ins using xG, which is more precise than just providing a number of shots, since each shot has a varied likelihood of scoring (as we understand the principle of Expected Goals). xG is assigned to the throw-in if the shot happens within 15 s (possibly within 30 s or 45 s) of the throw-in being taken.
- **A turnover** happens when the team loses the ball after the throw-in. **A quick turnover** occurs within seven seconds of the throw-in being taken.
- **A progressive throw-in** is one that moves the ball closer to the opposition by at least 20% of the distance between the throw-in's start position and opposition's goal or any other throw-ins aiming into the penalty box.
- **A flank** is defined as the edge of the pitch.

- **A penalty box** is a rectangular area extending approximately 16.5 mts on each side and 16.5 mts in front of the goal (International Football Association Board, 2021).
- **Passing combos** are combinations of one throw-in and two following passes.
- **A switch** happens when a player passes the ball straight from one flank to another (referred to as a **one-pass switch**) or when a player passes the ball from one side to the centre of the field and then targets another flank with the next pass (called a **two-pass switch**). **A throw-in with a switch** is a throw-in when the switch occurs within 7 s of the throw-in being taken.
- **Taker's goal** is always on the left-side of the horizontal pitch, and in the bottom of the vertical pitch in the visualisations. On the other hand, **Opposition's goal** is situated on the right-side of the horizontal pitch, and in the top of the vertical pitch in the visualisations. In chapter 4, the direction of the play will be illustrated by a gray arrow.
- **A backward throw-in** is a throw-in which is taken in the absolute value of an angle between 180 and 120 degrees.
- **A lateral throw-in** is a throw-in which is taken in the absolute value of an angle between 120 and 60 degrees.
- **A forward throw-in** is a throw-in which is taken in the absolute value of an angle between 60 and 0 degrees.
- **Passing (throwing) sonars,** was introduced by Eliot McKinley (McKinley, 2018)**,** displays frequency (via colours) and median distance of throw-ins within angle bins.
- **Bins** indicates intervals of certain continuous numbers (e.g., timing bins).
- **End distance from the opposition's goal** is the distance between the end location of the throw-in and opposition's goal, it is one of the features used to train the model.
- **Attacking zone** is the part of the pitch that surrounds the attacking third (that is closer to the opposition's goal).
- **Middle zone** is the part of the pitch that surrounds the attacking third.
- **Defensive zone** is the part of the pitch that surrounds the defensive third (that is closer to the taker's goal).
- **16 metres attacking/defensive zone** is defined as the part of the pitch that surrounds the area up to 16 mts from the opposition's / taker's goal; the definition is inspired by: (Stone, et al., 2021)

Figure 5: Throw-ins Glossary



Source: own elaboration

4 Practical Part

In this chapter, we analyse throw-in efficiency. In Section 4.1., we use Exploratory Data Analysis (EDA) to evaluate each feature individually. In Section 4.2, we construct gradient-boosted ensemble of decision trees to assess the global effect of all throw-in attributes on efficiency.

4.1 Exploratory Data Analysis (EDA)

Before building the model, it is essential to have an insight into the effect of numerous factors on the target variable. The features are used in the machine learning model based on the domain expertise in football and previously built models, e.g., Elliot's expected throw model (McKinley, 2018). In the following figures, we see how the features influence the target variable (completion), possession retention of throw-ins and the likelihood of scoring goals based on xG and Gls statistics, and also how they have been changed over time.

To begin, the probability of the throw-in going directly to a teammate, i.e., completion rate, is generally extremely high (0.892). This indicates that we have an unbalanced dataset for constructing a machine learning model in Section 4.3. Although the probability of possession being retained within at least seven seconds following the throw-in is less probable (0.578) than the completion rate. The chance of a successful throw-in with possession held is around 0.539, whereas the probability of a team retaining possession following a failed throw-in is extremely low (0.038). While the probability of a player throwing directly to a teammate and then losing control within seven seconds is around 0.353. The probability of scoring a goal from a throw-in within 15 s is approximately 0.004. In comparison with the probability of scoring from a throw-in, the mean xG after a throw-in is 0.003, indicating that teams are more efficient than it is expected. All these results are computed from the dataset derived from Wyscout data. Table 3 illustrates all mentioned rates.

Table 3: Probabilities of Types of Throw-in

Throw-in Type	Probability
Complete	0.892
Possession retained	0.578
Complete and possession retained	0.538
Incomplete and possession retained	0.038
Complete and lost possession	0.353
Scored goal after a throw-in	0.004
Mean xG after a throw-in	0.003

Source: own elaboration from Wyscout Data

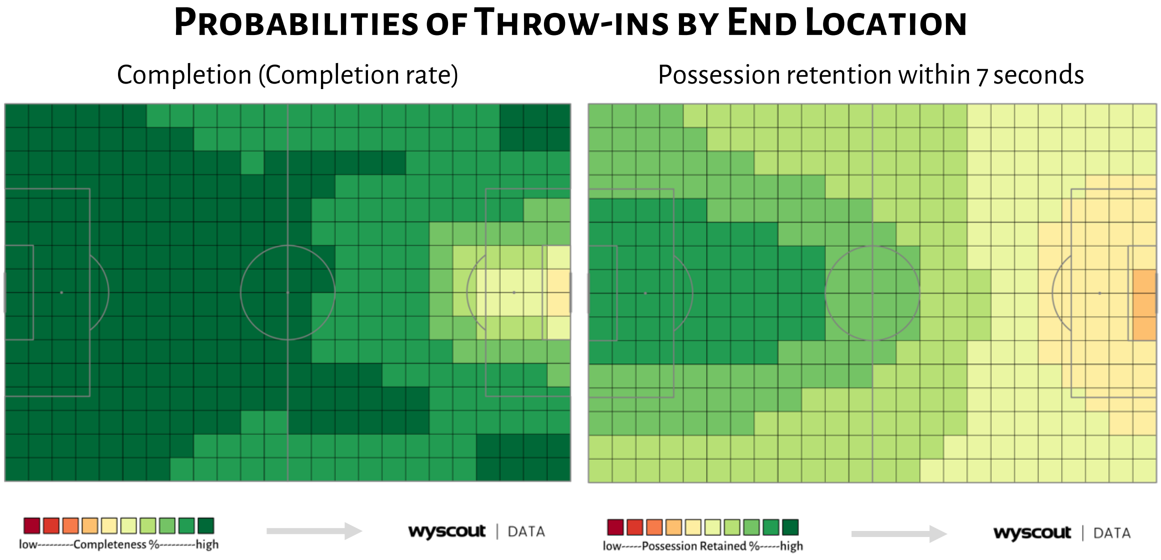
A few of the machine learning model's features are described through line plots with three distinct statistical indicators. These attributes include the following: throw-in distance (see Figure 9), throw-in angle (see Figure 11), as well as the time since the last action, i.e., timing (see Figure 12). Three distinct statistical indices are the target variable (completion rate), possession retention rate, and mean xGD. Each of the following figures illustrates a three-zone football pitch (defensive, middle, and attacking). Each zone has a plot corresponding to the observed values in that zone. Three distinct lines are used to denote each statistical indication on these graphs. The blue solid line indicates the completion rate value, the blue dashed line represents the likelihood of retaining possession, and the brown solid line implies the mean value of the xGD within 15 seconds for each bin in each zone.

4.1.1 Throw-in End Location

To describe the pitch visualisation in Figure 6, each bin has a value indicating the chance of the target variable being fulfilled. These values were smoothed to make visualizations more interpretable. A grey arrow situated next to football pitch visualization represents the direction of the play. The visualisation includes colorbars that describe the mathematical meaning of the colours. Heatmaps were normalized using minimum/maximum as 0/1 to emphasize patterns between these two statistical characteristics.

With regards to the end location of all throw-ins in the dataset, there is a general tendency that the closer the ball is thrown to the opposition's goal, the less likely a throw-in would be completed. Possession retaining is in a similar manner. Although, the higher chance of throw-in being possession retained is in the centre defensive area of the field from a goalkeeper’s perspective as Figure 6 demonstrates. However, tracking data containing the coordinates of all players on the field are not available to validate this idea, it seems that opposing players' pressing reduces the chance of possession being kept near the field's side-line.

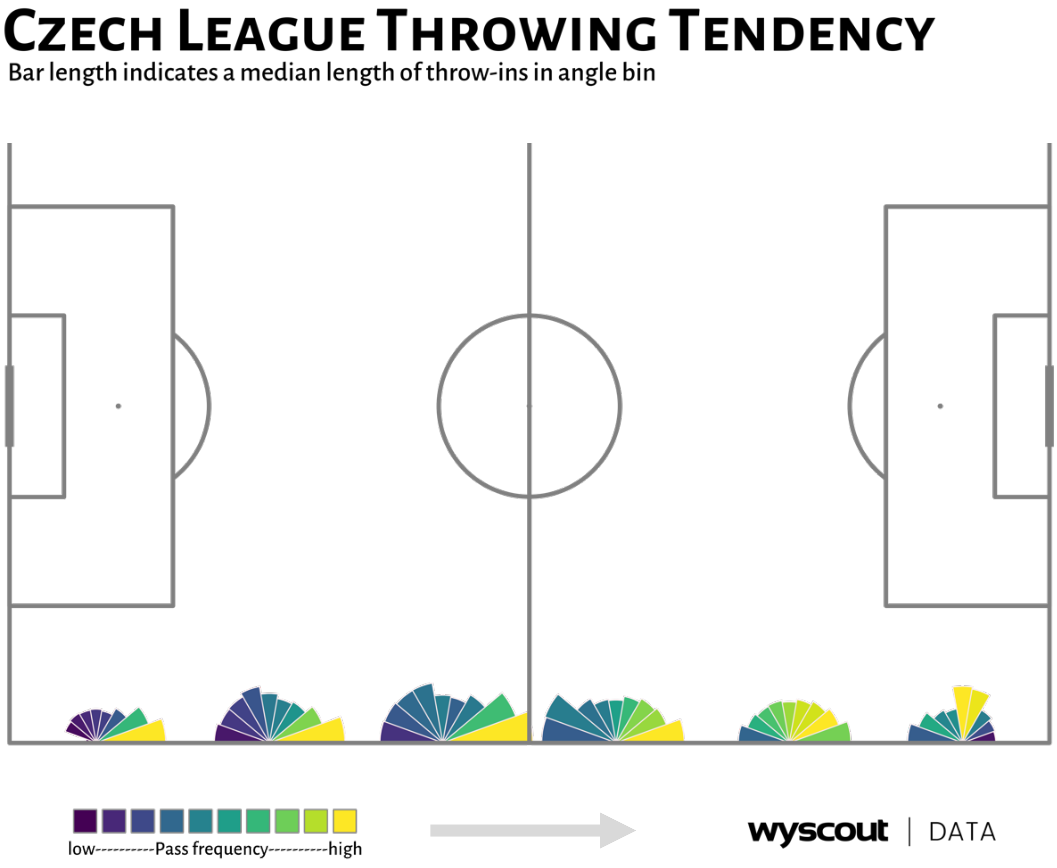
Figure 6: Completion and Possession Retention Rate of Throw-ins by the End Locations



Source: own elaboration from Wyscout Data

The pitch visualisation in Figure 7 depicts a football pitch with throwing sonars from certain start locations for both sides (situated in the bottom of the Figure 7), indicating the frequency and median length of throw-ins in each angle bin (direction). The longest average length across bins of successful throw-ins is 34.72 meters, while the shortest average length is 3.60 meters. As shown, the defensive and middle throwing sonars illustrate a preference for more progressive throw-ins in order to minimise the risk of losing possession closer to their own goal.

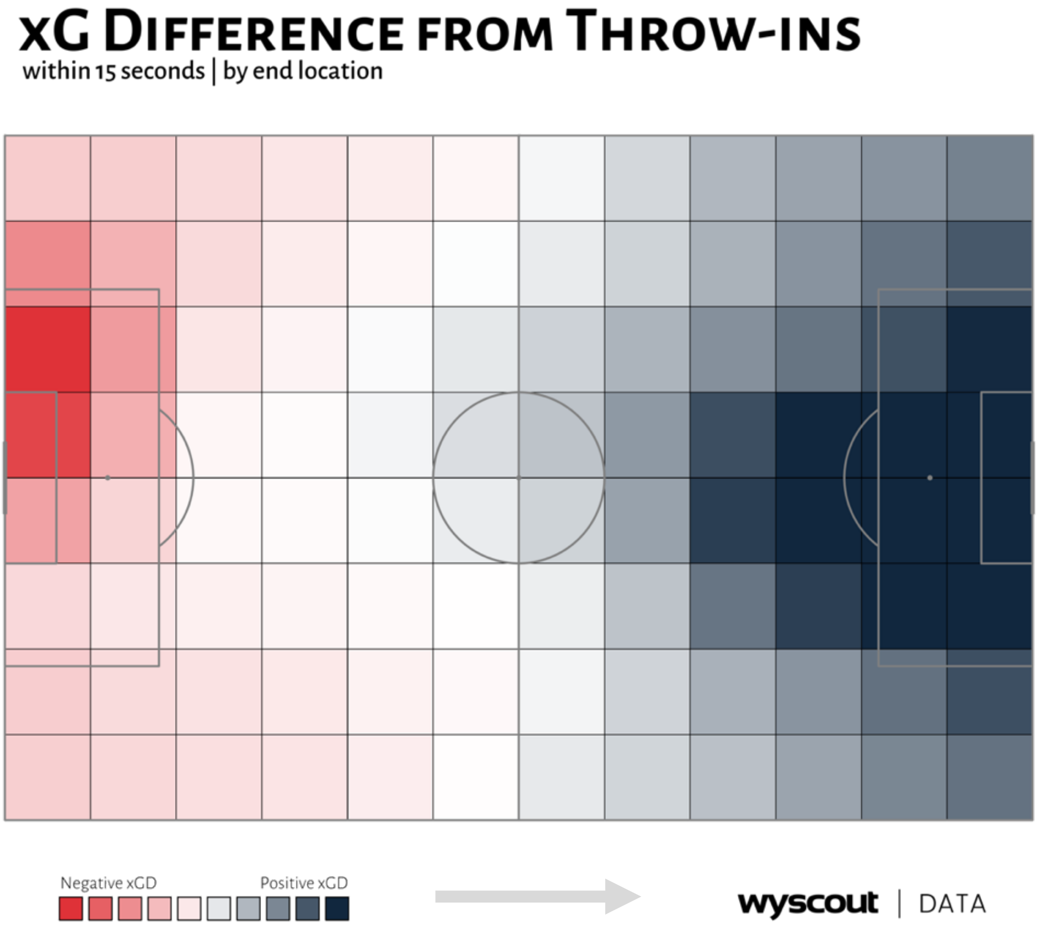
Figure 7: Throwing Sonars



Source: own elaboration from Wyscout Data

The essential aspect of this data analysis of throw-ins is to determine which area has the highest chance of scoring and conceding the goal (xG/xGA) following the throw-in within 15 seconds. Figure 8 represents difference of average xG and xGA values related to the final location of throw-ins. The negative xGD is shown by red bins, the equal xGD is indicated by white bins, and the positive xGD is expressed by blue bins. Not surprisingly, the xGD value is greater when the throw-in end location is closer to the opposition's goal, because this corresponds with xGD of all shots.

Figure 8: xGD from throw-ins by the end location

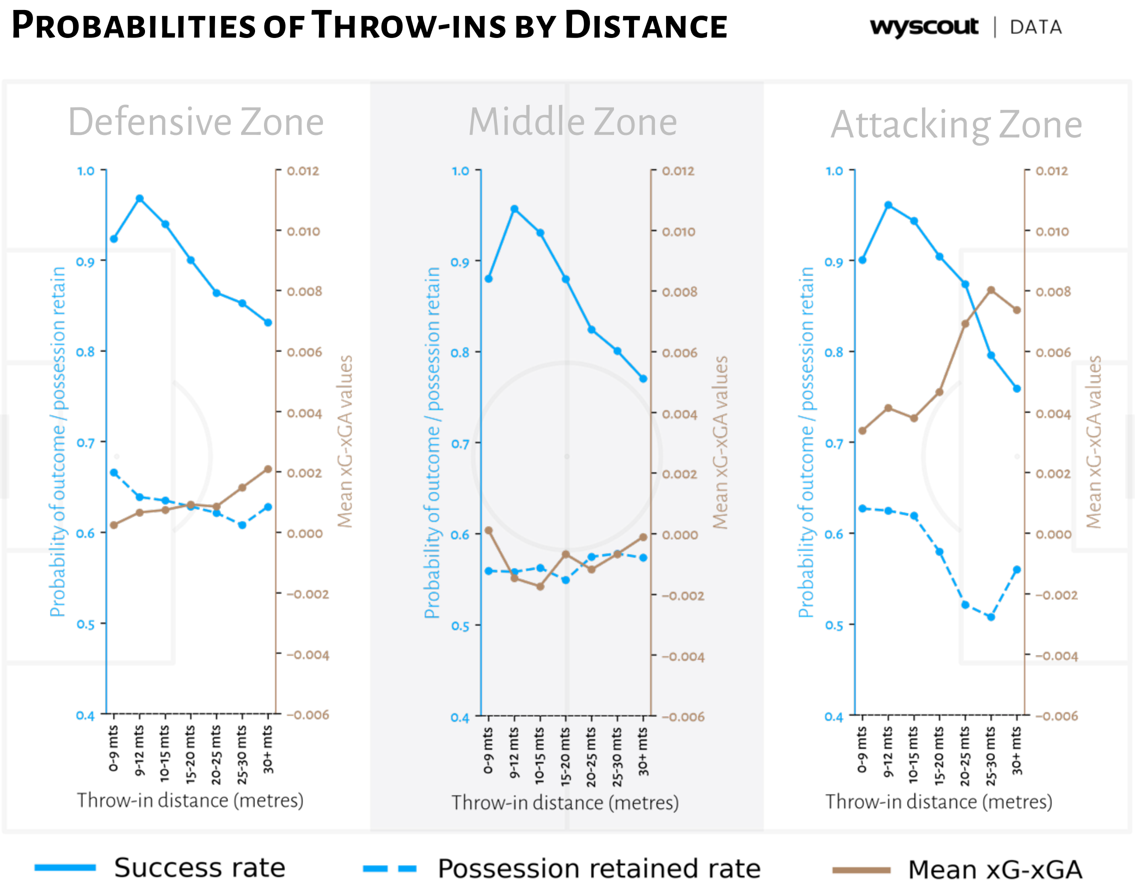


Source: own elaboration from Wyscout Data

4.1.2 Throw-in Distance

To begin, let us consider throw-in distance. The link between the specified statistical indicators and the distance of the throw-in inside each zone is visualized in Figure 9. The longer the throw-in, the lower the probability of success, which is valid for all zones. In the central zone of the pitch, the probability of the throw-in going directly to a teammate is greater. While the chance of completing is lower in the attacking zone. Ninety percent of throw-ins with a distance less than 5 mts are complete. The peak of the completion rate is between 5 and 10 mts. The completion rate of more than 30 meters throw-ins reduces to 78 percent.

Figure 9: Probabilities of throw-ins by the throw-in distance

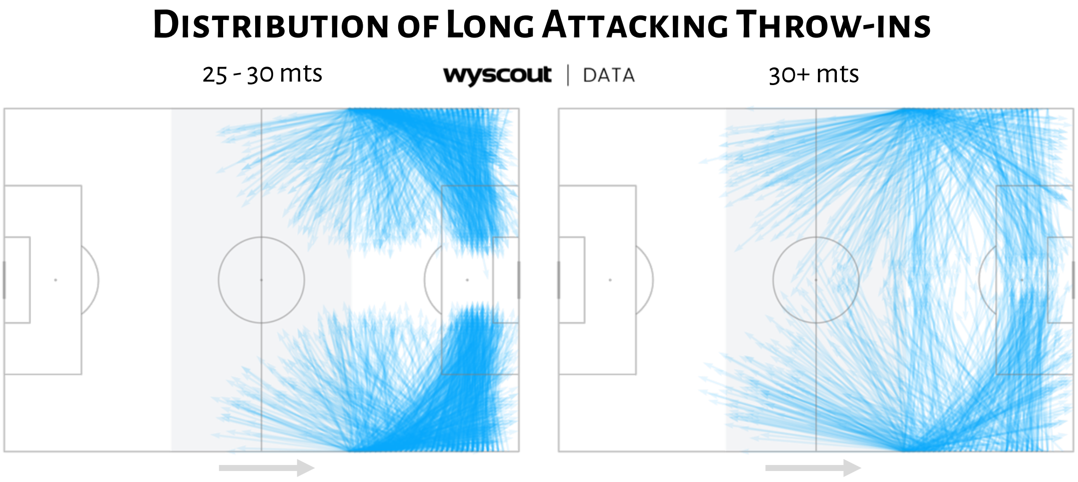


Source: own elaboration from Wyscout Data

In the defensive zone, possession retention is dropping as throw-in distance increases, except for the bin with 30+ mts throw-ins, when possession retention is slightly increasing. In the middle zone, possession kept within 7 seconds does not vary much among distance intervals; the likelihood fluctuates between 55 and 58 percent, which is a negligible change. Except for the 30+ mts bin for attacking throw-ins, where throw-in angles are more evenly distributed than in the 25-30 mts bin, we can observe the declining trend. In the middle zone in Figure 9, the average xGD is nearly constant. Moreover, there is a growing tendency within bins toward separating defensive and offensive zones.

To illustrate, Figure 10 displays the distribution of long attacking throw-ins for these two previously mentioned bins, with blue arrows indicating the direction of each throw-in. Three distinct zones are denoted by the gray shade. Both of these bins offer a single pitch to visualize their distribution of throw-ins. The bin with a throw-in length of 25-30 mts contains a higher ratio of progressive throw-ins compared to the bin with a throw-in length of 30+ mts. We already know that, according to the xG model, throwing the ball closer to the opposition's goal increases the probability of scoring a goal (see Figure 8). Based on the fact presented in Figure 6, the bin with a length of 30+ mts has a higher probability of possession being kept and has a lower mean xGD than bins with a length of 25-30 mts.

Figure 10: The distribution of long attacking throw-ins

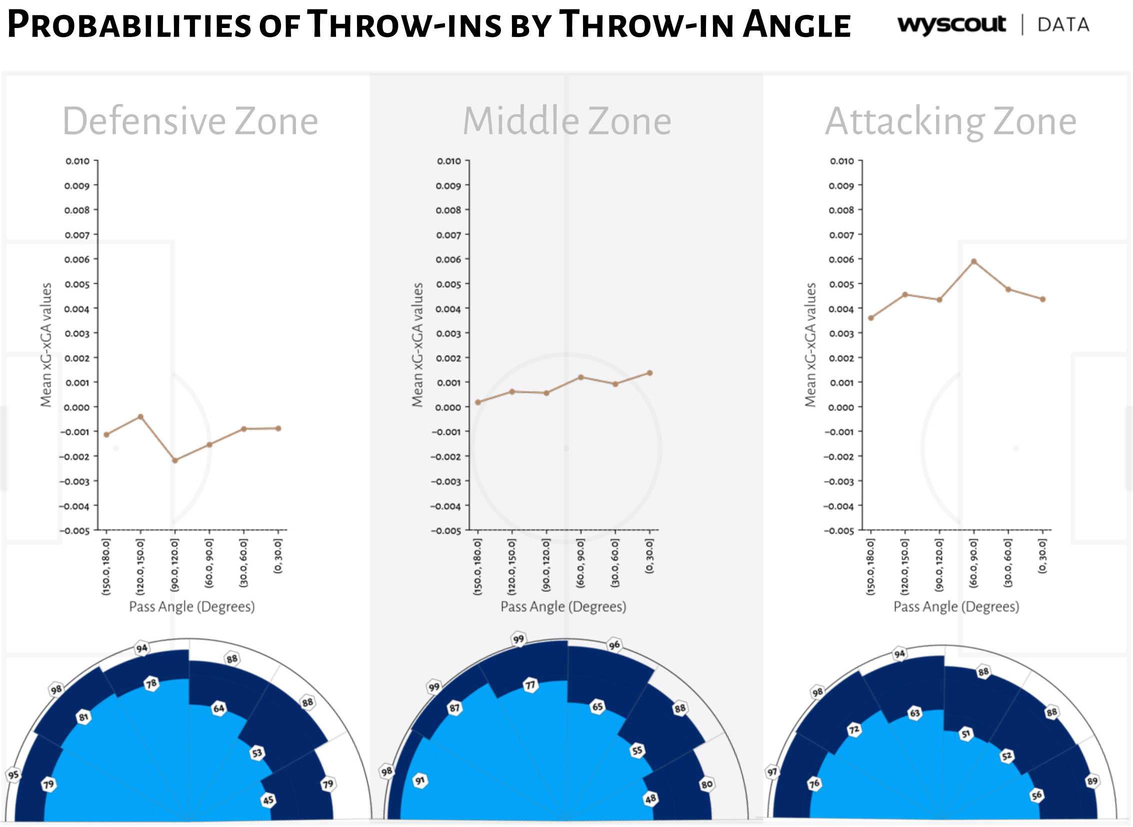


Source: own elaboration from Wyscout Data

4.1.3 Throw-in Angle

Next, we consider throw-in angle. Throwing sonars are extra components to the Figure 11, rather than line plots like in the preceding probability chart Figure 9. Each pizza slice represents a bin. The sequence of the bins begins on the left with the bin with an angle of 180-150 degrees and progresses in steps of 30°, as shown in the line plot above. The dark blue pizza slice reflects the completion rate of the bin's throw-in in the zone, whereas the light blue slice indicates the possession retention rate. The values in hexagons on the slice ranging from 0 to 100 denote their percentage. As throwing sonars show in Figure 11, there is no significant (principial) difference between each zone. Throw-ins directed toward the opposition's goal have a lower probability of being completed or retained.

Figure 11: Probabilities of throw-ins by the throw-in angle with sonars



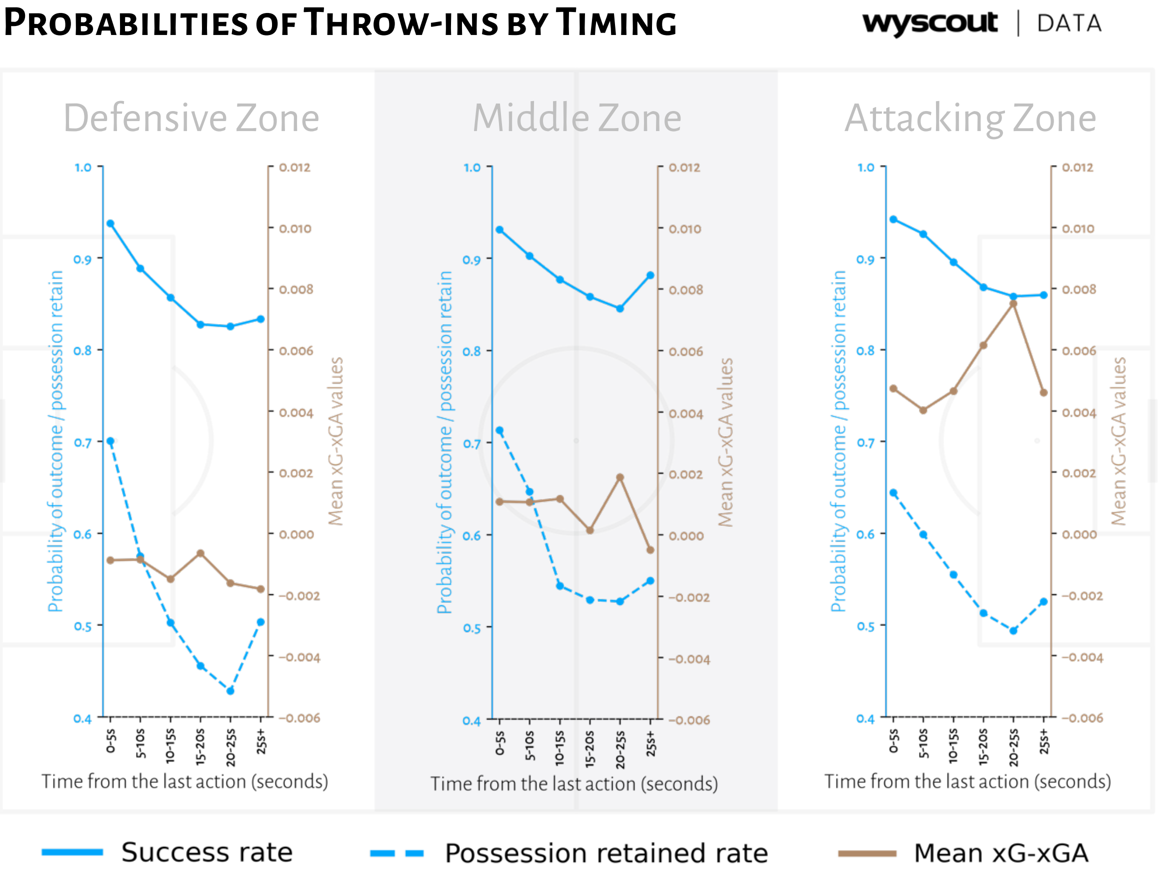
Source: own elaboration from Wyscout Data

The fact that the angle bin with the 120º–150º is closer to the positive value of the mean xGD in the defensive zone is an interesting factor to consider. This is also the angle bin with the greatest likelihood of taking a completed or retained throw-in. The mean of xGD of an angle bin with perpendicular and slightly progressive throw-ins lowers to a local minimum. Only then, the two most progressive angle bins (with range of 0º to 60º) begin to get closer to the positive value of the mean xGD. In other words, the most successful strategy is to initiate throw-ins from the defensive zone slightly backward. In the middle zone, a little upward trend over bins suggests that taking more progressive throw-ins slightly increases the likelihood of scoring and avoiding a conceded goal. The rising trend of xGD is valid for the attacking zone. It's worth noting that when a taker, standing in the attacking zone, gets closer to the flag, the xG (as well as success and possession retention rates) trend reverses, making it more advantageous to take perpendicular or slightly backwards throw-ins in terms of throw-in angle probability link.

4.1.4 Timing of Taking Throw-ins

Figure 12 shows the probability associated with the characteristic – time since the last action. In all zones, the faster the taker throws, the greater the possibility of success or possession retention, however the completion rate begins to decline after bins with 10-15 seconds after the last action. Big jumps across bins for possession retention tells us that timing of a throw-in has bigger impact on possession retention than completeness. Possession retention, on the other hand, operates similarly, but with a little delay of around 5 s. Which we may attribute to domain knowledge, as the opponent is capable of structuring a defence within 15 s, making a successful (or possession retained) throw-in more difficult to execute.

Figure 12: Probabilities of throw-ins by timing



Source: own elaboration from Wyscout Data

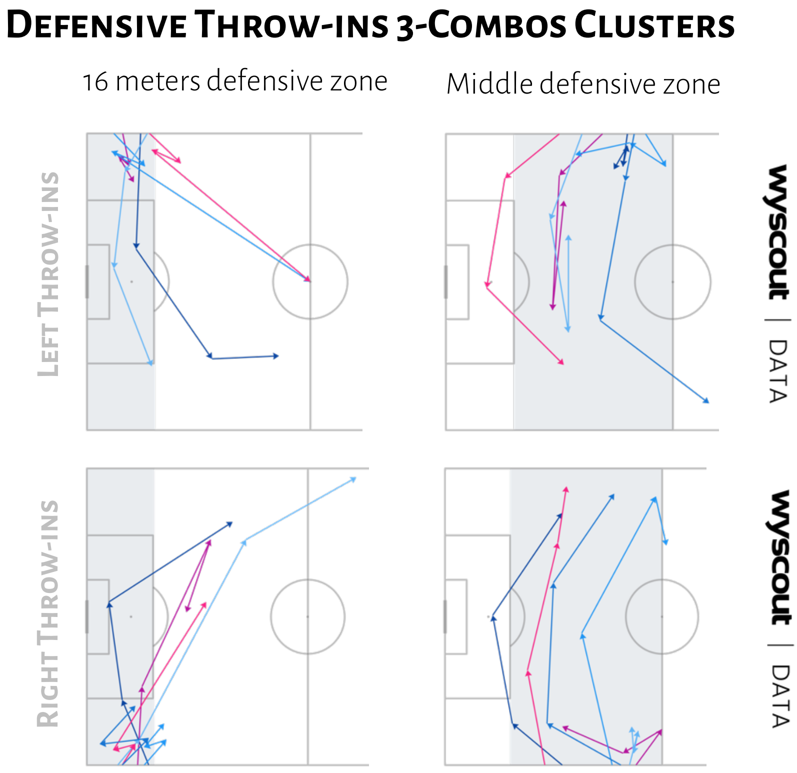
Defensive and middle zones are comparable in terms of the mean xGD. Figure 12 indicates that timing does not significantly influence xG from throw-ins originated in these two zones. In the attacking zone, the mean xGD seems to be greater when the taker waits for an ideal option such as long throw-ins aimed towards the penalty box, which require more time since the taker must run up to produce greater distance and occasionally clean the ball for a better grip. As previously stated, based on the average of xGD values, this kind of throw-in has a better likelihood of scoring than conceding.

4.1.5 Throw-ins Combos

Another significant factor influencing possession retention is a few initial successful passes following throw-ins, which may be analysed by clustering those pass combinations (also known as combos). We selected throw-ins when possession is kept for at least 10 s following the throw-in. The OPTICS clustering approach is used for clustering of passing combinations based on following characteristics: start and end location of each pass in combos, angle between those passes, and distance of those passes. To obtain more realistic results, the clustering process is used eight times with different throw-in start location zones. To aid with visual interpretation, these clustering combinations include a throw-in and just two subsequent successful passes in the following figures. The top 5 clusters for each zone are displayed, and each individual cluster of combination is represented by a different colour, as seen in Figure 13 and 14. Clusters are visualized as the median arrow of all arrows that belong to the cluster.

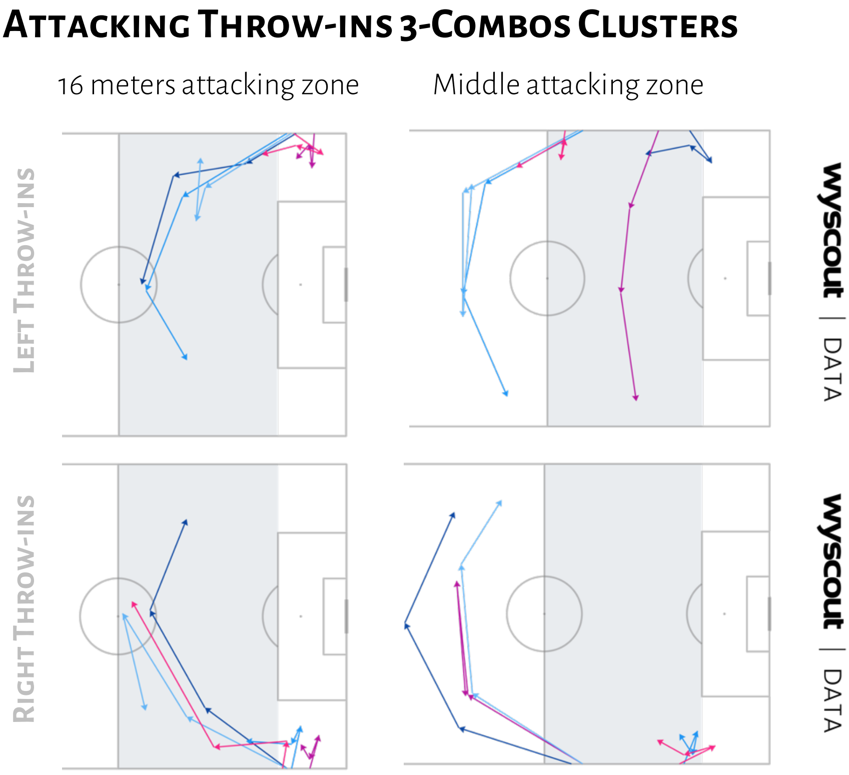
The clusters discovered for the defensive zone demonstrate one dominant strategy for keeping possession of the ball. This strategy involves throwing the ball around 10 mts away, passing backward, and then switching in one pass. Additionally, we can see that the team follows the throw-in with a two-pass switch rather than a backward and one-pass switch (see Figure 13). Cluster analysis of the attacking zone reveals comparable strategies (to the defensive zone) for retaining possession of the ball (see Figure 14). The first method includes throwing the ball inside a 7-meter radius, followed by another short-pass combo. The second approach contains a longer backward throw-in followed by a two-pass switch aiming towards another flank. In the sixteen metres attacking zone in contrast to other zones, the dominating combination consists of longer passes aiming to the centre of the pitch from the goalkeeper's perspective following a pass back to the same side where the ball was received.

Figure 13: Defensive throw-ins combos



Source: own elaboration from Wyscout Data

Figure 14: Attacking throw-ins combos



Source: own elaboration from Wyscout Data

4.1.6 Throw-ins with a Switch

After gaining insight of the patterns on combos, we know that switching from one side to the other is a part of passing combinations for retaining possession for at least 10 s. Following that, it is also useful to investigate how switches influence scoring and conceding probability. To assess this issue, only switches made by the team that takes a throw-in within 7 seconds after the throw-in is counted. Figure 15 illustrates how the likelihood of scoring or conceding a goal changes if the team chooses to play a switch after the throw-in within 15, 30, or 45 s. Blue-coloured lines symbolize xG, whereas red-coloured lines denote xGA. Each frame of time has a different type of line; a solid line indicates mean xG/xGA within 15 seconds, a dashed line shows xG/xGA within 30 s, and a dashed-dot line implies xG/xGA within 45 s after the throw-in.

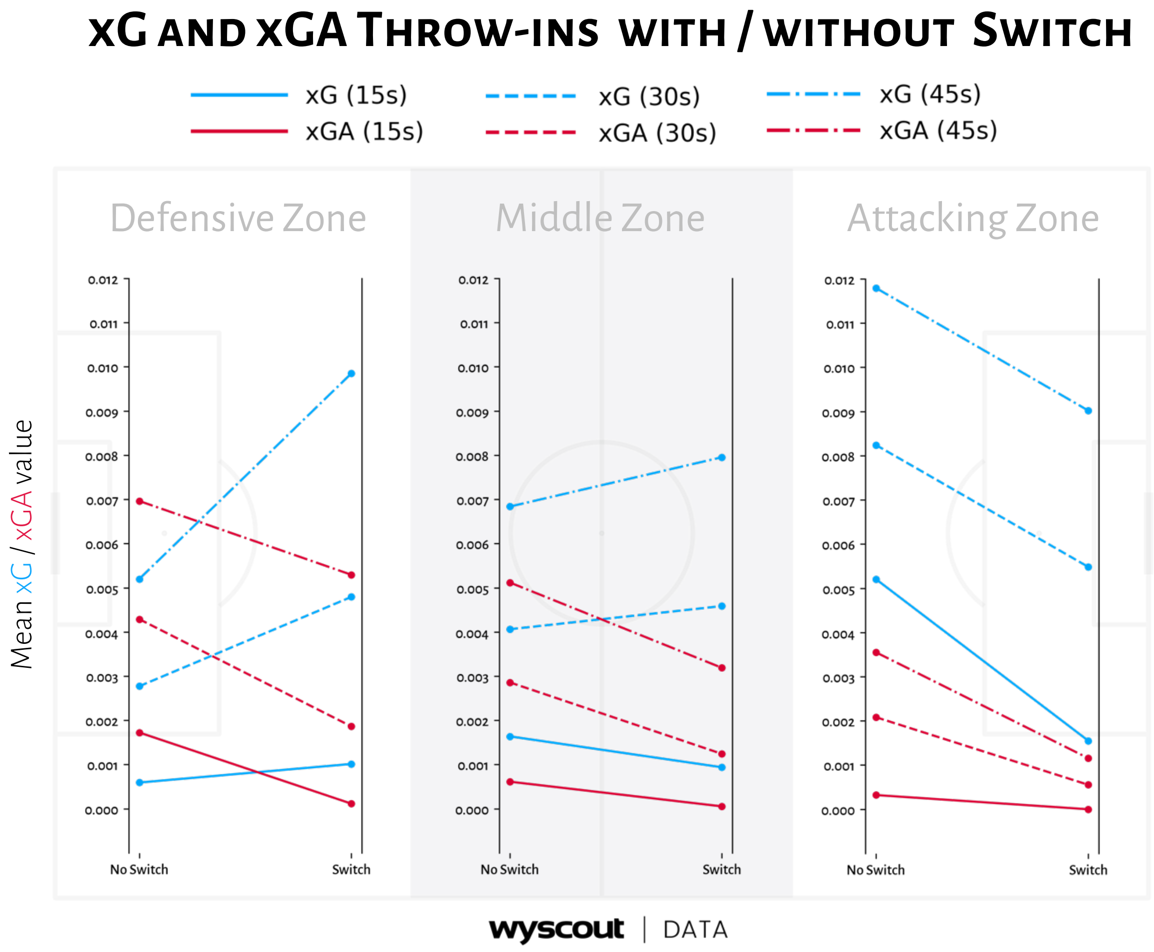
In the defensive zone, switches often improve the probability of scoring while decreasing the danger of allowing a goal. Despite the fact that the defensive zone is logically the furthest from the opposition goal, the positive change in mean xG value within 15 s between no switch and a switch is small. It is significant that it reduces the chance of conceding a lot. Additionally, the mean value of xG and xGA within 45 s shows a big positive effect from no switch to a switch, indicating that switching after throw-ins in the defensive zone is effective.

While there is a minor reduction in the likelihood of conceding within 15 s in the middle zone, there is also a reduction in the likelihood of scoring a goal, which is influenced by a less progressive type of play, such as switches. Switches have a positive effect on the mean of an xG value within 30 s and 45 s, indicating that switches in the middle zone are just as effective at reducing conceded goals and increasing the chances of scoring a goal as they are in the defensive zone.

The mean values of xG and xGA in the attacking zone after a throw-in with switches are diametrically opposite to those created and allowed in the defensive and middle zones within 7 seconds. When a team plays a switch, the xG and xGA values significantly drop within any given time (15, 30, 45 s). In other words, switches have a negative influence on

GIs scoring and GIs conceding, suggesting that switches after a throw-in in the attacking zone are not as effective as switches in the other two zones.

Figure 15: xGD with / without a switch after a throw-in

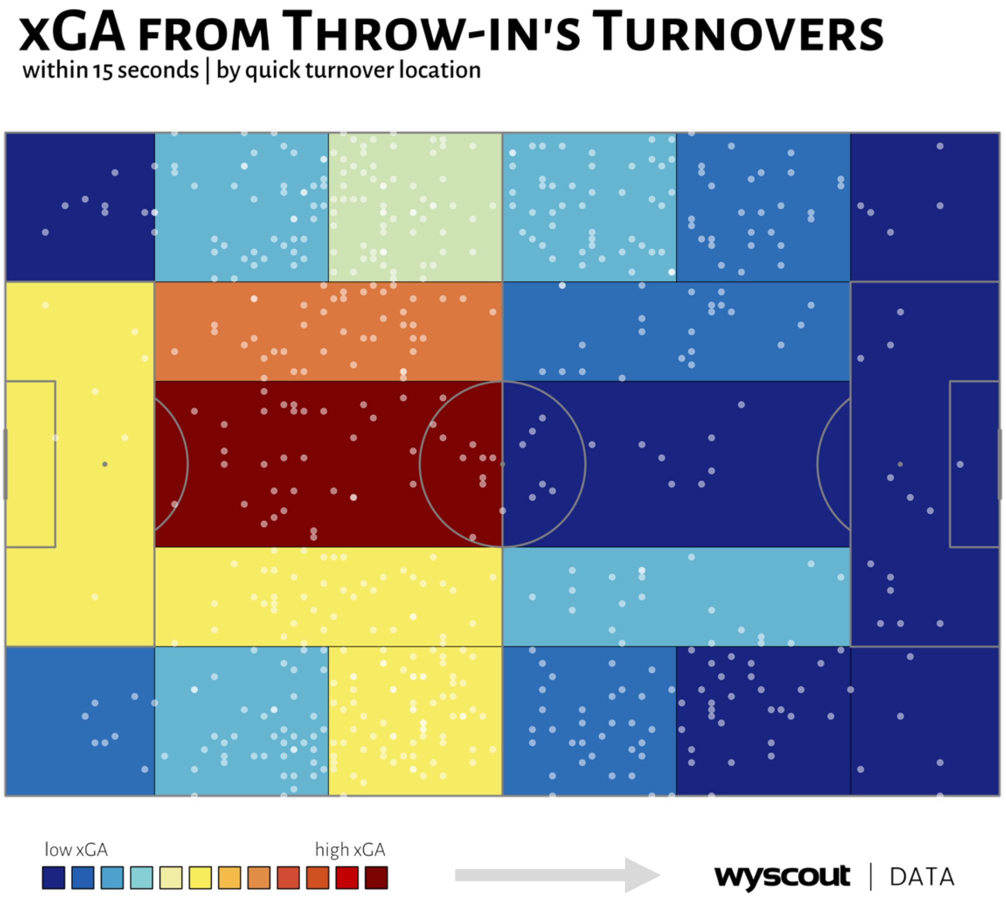


Source: own elaboration from Wyscout Data

The relationship between the chance of conceding a goal and the location of quick turnovers (a loss of possession within 7 s) after the throw-in is shown in Figure 16. Each mark on the pitch corresponds to the position of a quick turnover. The dark red area denotes the zone with a high probability of conceding, whereas the dark blue rectangle represents the opposite. According to Figure 16, if the first pass of a two-pass switch is incomplete and results in the loss of possession, there is a greater likelihood of conceding a goal (see the dark red rectangle in the defensive part of the field in Figure 16), implying that if the circumstances allow for a one-pass switch from one flank to the other, it is supposed to be a safe choice in contrast to the loss the ball in the centre of the field from the goalkeeper's perspective.

Additionally, Figure 16 indicates that the safer option (in terms of xGA) is the most often used throw-in tactic in the defensive and middle zone among Czech league coaches (see the density of scatter points in Figure 16): long throw-ins aiming towards tall strikers. Regardless, this option reduces the chances on successfully finding a teammate and retaining possession, which means increasing the likelihood of the opponent's creating a chance to score. To maintain control, it is critical to seek open space rather than just throwing the ball to a striker and leaving him to battle for it, which gives him a 58% chance of winning the duel, according to the dataset from Wyscout data (Wyscout, 2022), which is nearly equivalent to the probability of flipping a coin.

Figure 16: xGA from Throw-in's Turnovers

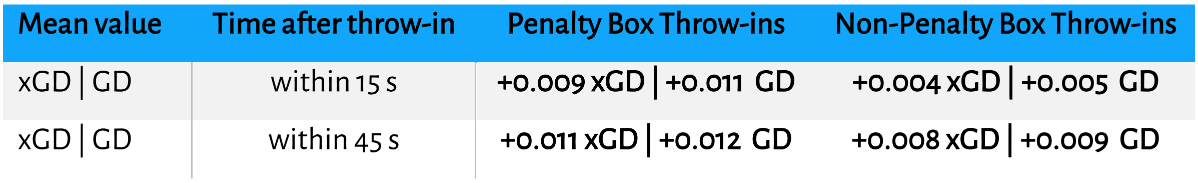


Source: own elaboration from Wyscout Data

4.1.7 Attacking Throw-ins

As we have a better understanding of throw-ins, the next topic are attacking throw-ins aimed towards the penalty box, which appears to be a potential weapon for increasing chances on scoring rather than conceding a goal, in contrast to attacking throw-ins targeted outside of the penalty box. To compare these two types of throw-ins, their **mean xGD** and **goals difference (GD)** were computed. The mean of an xGD is calculated by subtracting the mean of the xG created by the throwing team from the mean of the xG generated by the opposing team from their throw-ins. Similarly, the mean of the GD is calculated. Table 4 demonstrates that it is more efficient to play throw-ins that end inside the penalty box, since they have a higher probability of scoring and not conceding a goal, as determined by the xGD and GD statistics. In Table 4, the xG and Gls statistics are counted if the shot is taken after a throw-in within 15 s or 45 s.

Table 4: Attacking throw-ins xGD and GD comparison table



Mean value	Time after throw-in	Penalty Box Throw-ins	Non-Penalty Box Throw-ins
xGD | GD	within 15 s	+0.009 xGD | +0.011 GD	+0.004 xGD | +0.005 GD
xGD | GD	within 45 s	+0.011 xGD | +0.012 GD	+0.008 xGD | +0.009 GD

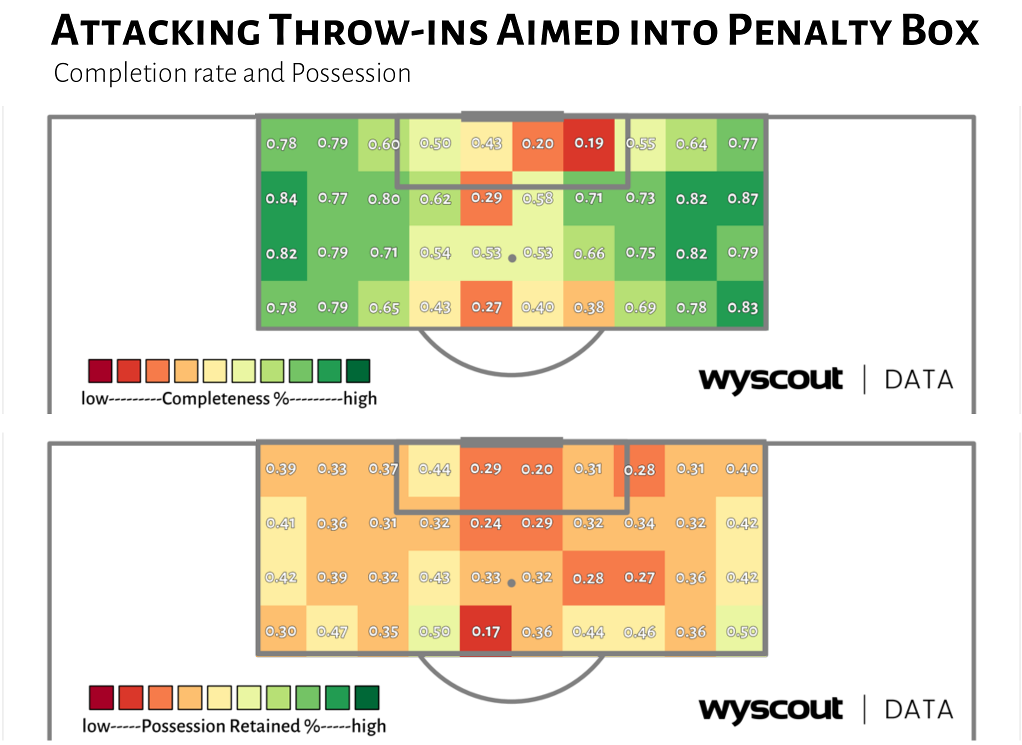
Source: own elaboration from Wyscout Data wyscout | DATA

In Figure 17, two football field visualisations with an attacking penalty box are shown. These graphs illustrate the relationship between the end position of offensive throw-ins that enter the penalty box and their bins' completeness rate, possession retention rate. To explain the results, there is a relatively high expectancy of being accurate with throw-ins aimed towards the penalty box's half-spaces and near-post throw-ins. Possession retention within seven seconds seems to be less likely to be held the closer the opponent goal is compared to completion rate.

The difference of timing efficiency between attacking throw-ins targeting the penalty box and those that finish outside the penalty box are described by the xG metric within 15 seconds in Figure 18. The solid line represents attacking throw-ins aiming into the box, whereas the dashed line symbolizes the attacking throw-ins playing outside of the box.

Figure 18 demonstrates that the bin with a time between 20-25 s after prior action has the greatest xG and xGD.

Figure 17: Probabilities of attacking throw-ins by the end location

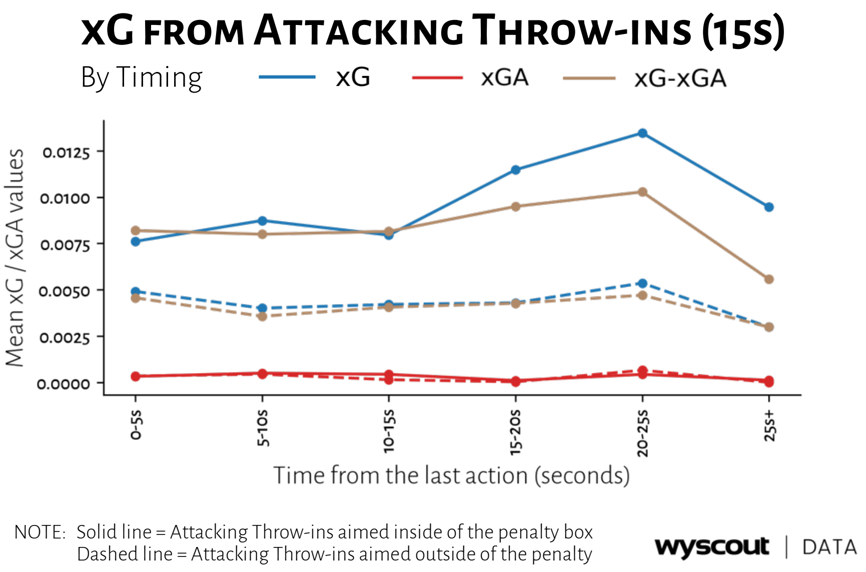


0.78	0.79	0.60	0.50	0.43	0.20	0.19	0.55	0.64	0.77
0.84	0.77	0.80	0.62	0.29	0.58	0.71	0.73	0.82	0.87
0.82	0.79	0.71	0.54	0.53	0.53	0.66	0.75	0.82	0.79
0.78	0.79	0.65	0.43	0.27	0.40	0.38	0.69	0.78	0.83

0.39	0.33	0.37	0.44	0.29	0.20	0.31	0.28	0.31	0.40
0.41	0.36	0.31	0.32	0.24	0.29	0.32	0.34	0.32	0.42
0.42	0.39	0.32	0.43	0.33	0.32	0.28	0.27	0.36	0.42
0.30	0.47	0.35	0.50	0.17	0.36	0.44	0.46	0.36	0.50

Source: own elaboration from Wyscout Data

Figure 18: xGD from attacking throw-ins by timing



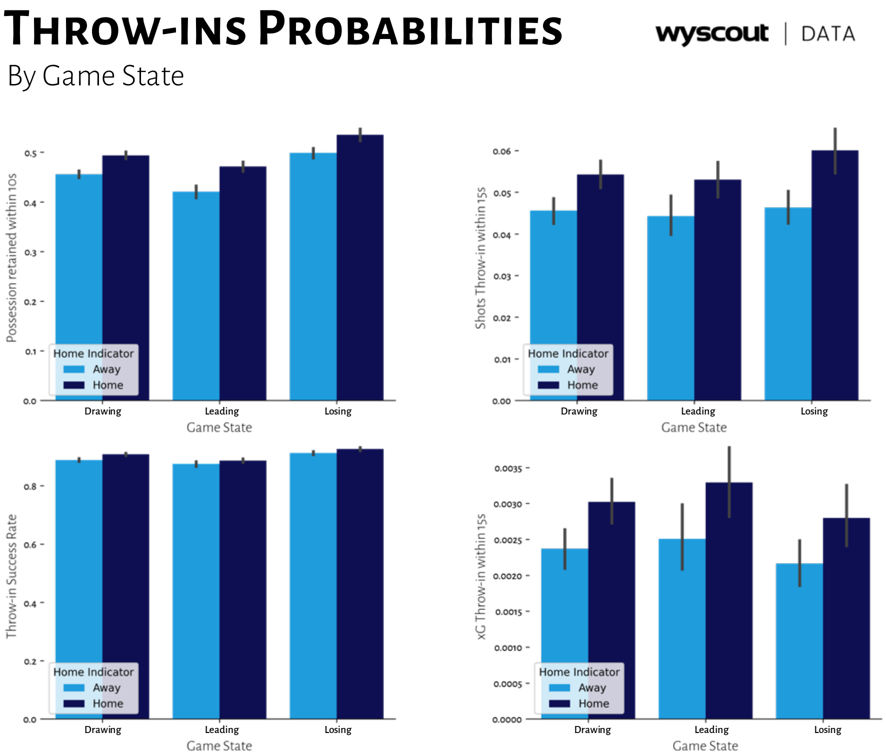
NOTE: Solid line = Attacking Throw-ins aimed inside of the penalty box
Dashed line = Attacking Throw-ins aimed outside of the penalty

Source: own elaboration from Wyscout Data

4.1.8 Throw-ins terms of Game State

The term "game state" refers to the score between two teams during a match. We classify it as being in one of three states: leading (with more goals than the opponent), drawing (level), or losing (less goals than the opponent). The probability for each game state is shown in Figure 19, including completeness, possession retention, shooting from the throw-in, and expected goals created from throw-ins. In Figure 19, we have added a hue parameter labelled the "home indicator," which illustrates how the likelihood changes depending on whether the team is playing at home or away. A gray line on each bar symbolizes 95% confidence interval. As seen, the game situation has an effect on the efficiency of throw-ins. Apparently, teams, that are losing, seem to be more effective with throw-ins in all areas except xG, which is closely connected with overall team performance.

Figure 19: Probabilities of throw-ins by the game state

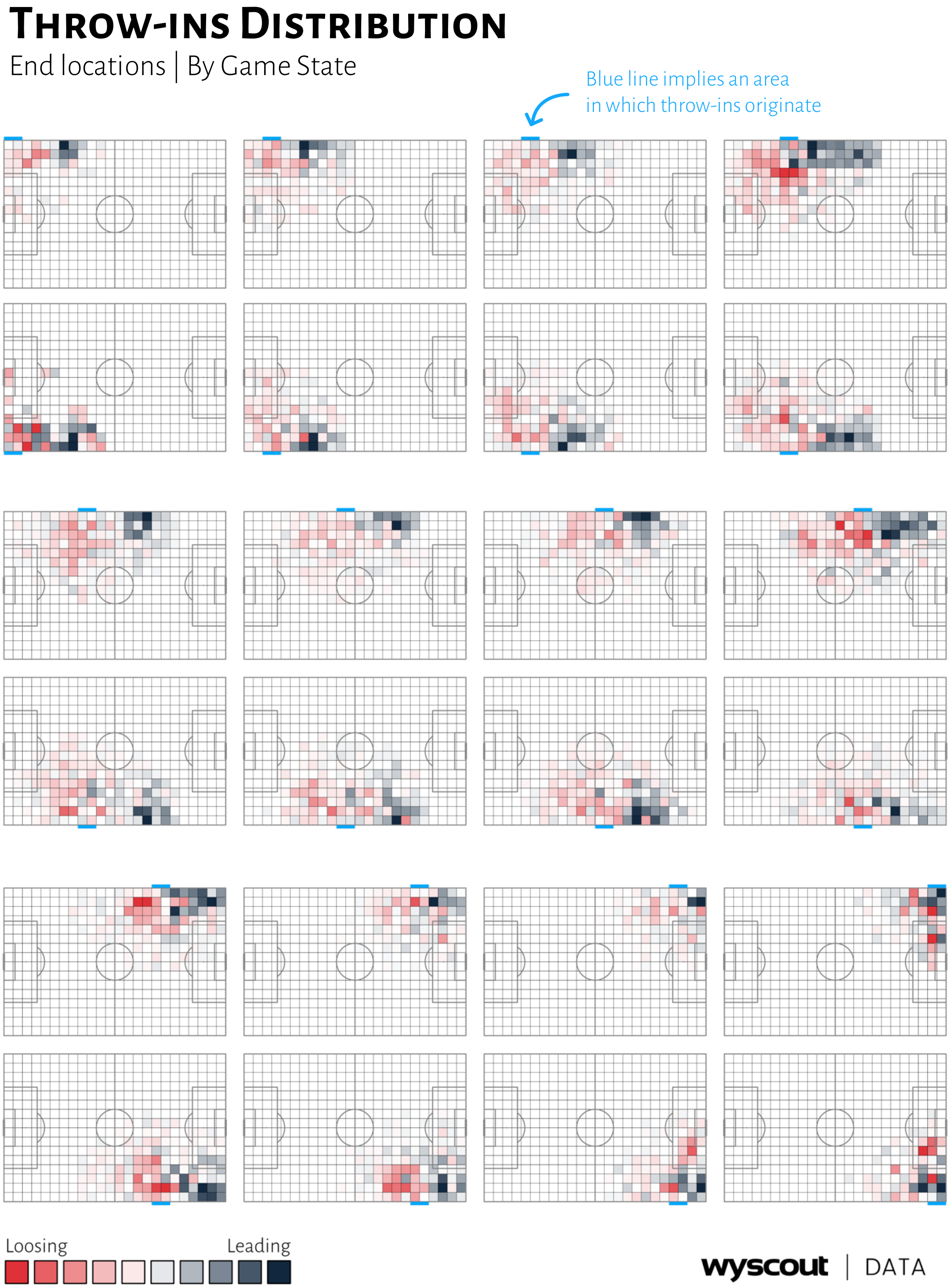


Source: own elaboration from Wyscout Data

As a result, we examined the distribution of when the team is leading and when it is losing, as well as the areas where the team is more efficient at generating opportunities and controlling the ball. The study, which was conducted by subtracting two normalised distributions, demonstrates where the distribution of throw-ins while the team is winning (blue) and when the team is losing (red) is more frequent, with white representing the balance between the distributions (see Figure 20). The blue line outside of each pitch implies an area in which throw-ins originate.

The visualisation demonstrates that when a team is losing (red colour), they tend to throw at a less risky angle and look for positions that are generally more advantageous for ball possession, such as throw-ins aimed more towards the central zone of the pitch. The exception of attacking throw-ins, where the losing team attempts to use more long throw-ins aimed into the penalty box to create more chances, which is more effective, as we already know (see Table 4). This demonstrates that the leading team often surrenders in order to pursue a more successful plan.

Figure 20: The distribution of throw-ins by the game state

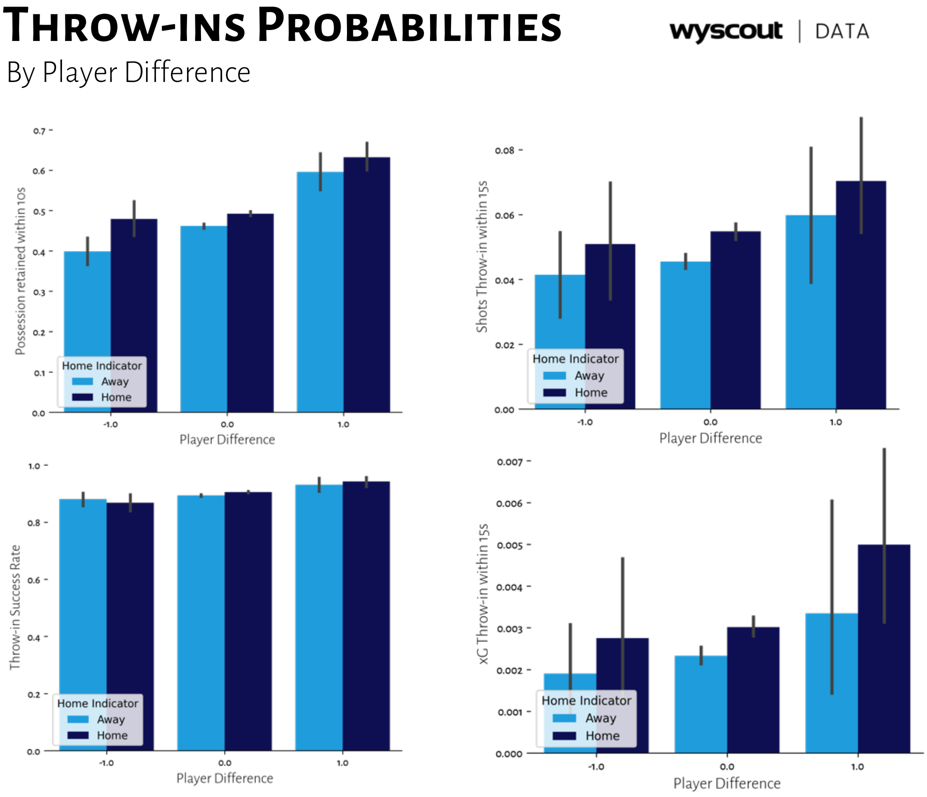


Source: own elaboration from Wyscout Data

4.1.9 Throw-ins terms of Player Difference

Another aspect that may affect the model is the player difference, in other words, the team that throws plays with more or less players due to player exclusions. As can be seen in Figure 21, the player difference affects the effectiveness of throw-ins. Thus, teams with an additional player are more successful with throw-ins in all areas, as well as in terms of xG created.

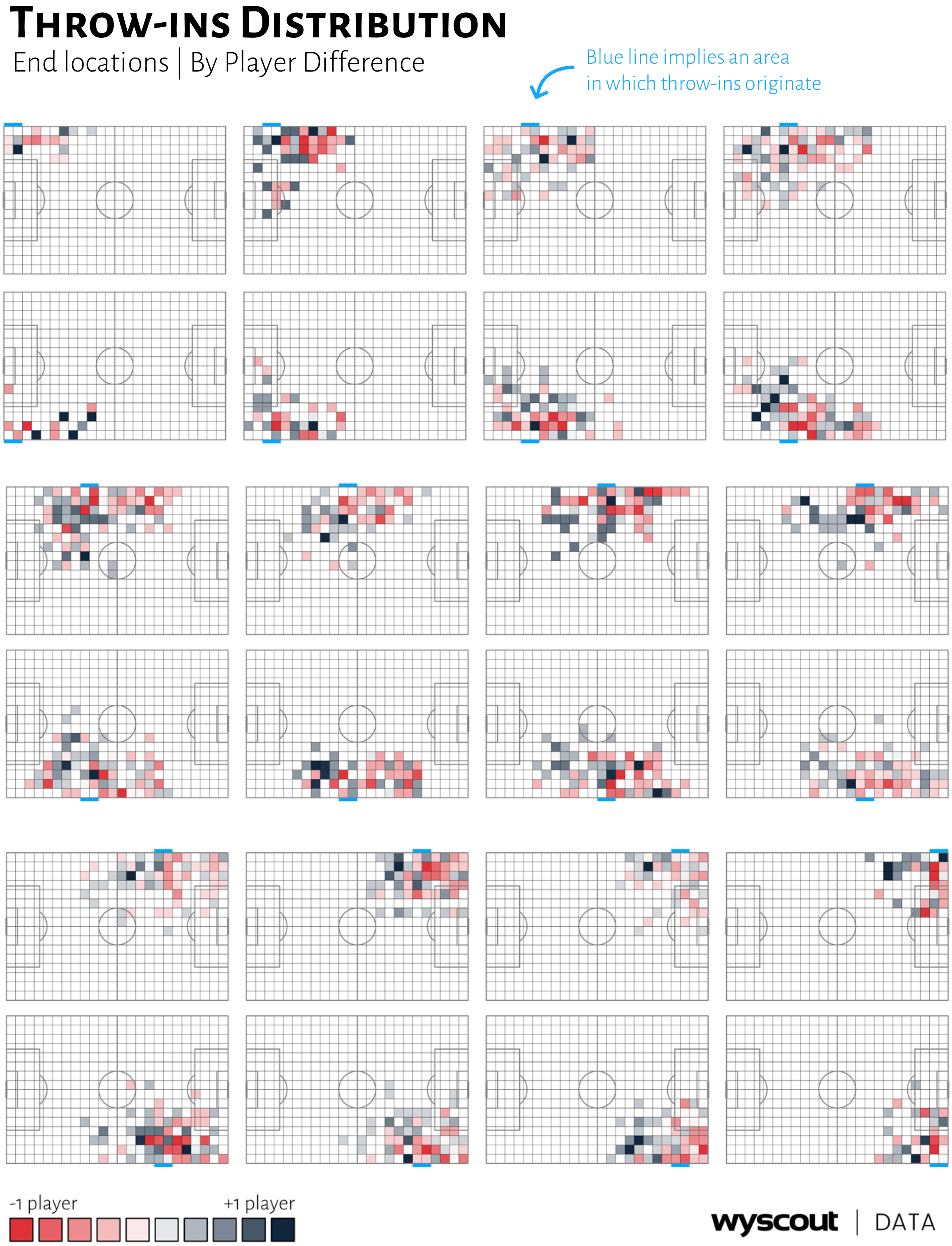
Figure 21: Probabilities of throw-ins by the player difference



Source: own elaboration from Wyscout Data

On the same premise, we investigated the distribution of throw-ins in which zones the team with one more player on the field (blue) is more common than the team with one less player on the pitch (red). As seen in Figure 22, there are no obvious patterns in terms of throw-in end location in contrast to throw-in distribution in terms of losing vs leading, but we can conclude that teams with one less player use more progressive throw-ins. When we examine the team's data set with one extra player, we discover that 32% of their throw-ins are progressive, however when we study the team's data set with one less player, we discover that 45% of their throw-ins are progressive. Despite this fact, the teams with one less player are not efficient at creating chances in terms of number shots and their xG values.

Figure 22: The distribution of throw-ins by the player difference



Source: own elaboration from Wyscout Data

4.1.10 Feature Analysis Summary

After our investigation of several factors affecting throw-ins, we conclude what data analysis of throw-ins reveals. Table 5 shows the best strategy of taking a throw-in in certain zone, based on the data we work with. For example, the ideal option how to take a throw-in in the attacking zone to create chance is to play a long (distance bin = 25-30 mts), slow (timing bin = 20-25 s), and slightly backward (angle bin = 90-60 °) throw-in without utilising a switch, with the throw-in aimed towards the penalty box. Some cells include stars after the value, indicating that although there is no evident pattern, that value, is the best choice. Additionally, some of them are empty, indicating that they have not been examined.

Table 5: Feature Summary

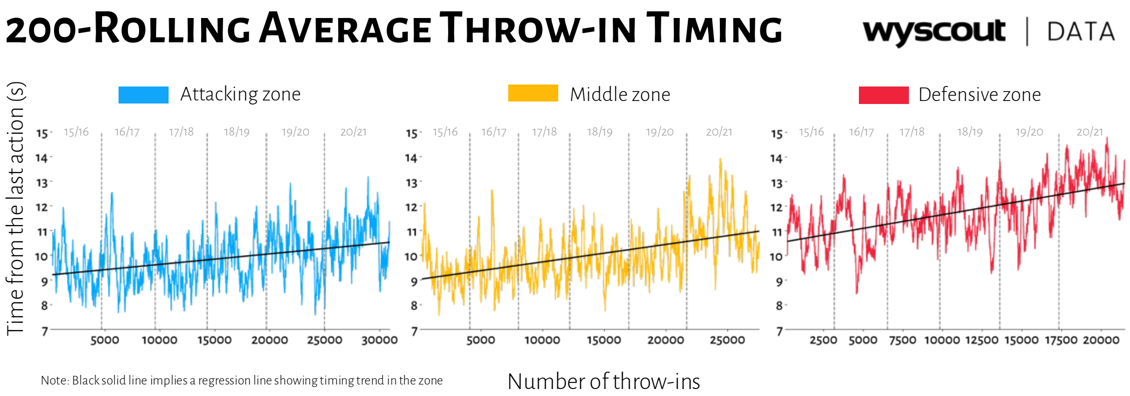
Throw-in Origin	Purpose	Distance	Timing	Angle	Switch	Box
Defensive zone	Completion	9-12 mts	0-5 s	120-150°		
	Possession Retention	0-9 mts	0-5 s	120-150°		
	Creating Chances (xGD)	30+ mts	15-20 s*	120-150°	YES	
Middle zone	Completion	9-12 mts	0-5 s	90-150°		
	Possession Retention	25-30 mts*	0-5 s	150-180°		
	Creating Chances (xGD)	0-9 mts	20-25 s*	0-30°*	YES	
Attacking zone	Completion	9-12 mts	0-5 s	120-150°		RATHER NO
	Possession Retention	0-9 mts	0-5 s	180-150°		NO
	Creating Chances (xGD)	25-30 mts	20-25 s	90-60°	NO	YES

Source: own elaboration from Wyscout Data

We also examine trends over time and compare them to the results of data analysis, which we observed in previous chapters. This provides insight into how the Czech Football League responds to the effectiveness of taking throw-ins. Two line-plots depict patterns in the timing and distance of throw-ins, which was separated into three unique zones to aid in gaining insight into trends. These two line-plots illustrate a regression line that makes it easier to visualize the trend of changing data (“time since the last action” or “throw-in distance”). Additionally, we use a rolling average of 200 throw-ins to smooth the line. As seen in the following figures, gray dashed lines mark the boundaries of each season, and their start and finish dates are noted in the same colour at the top of each zone's plots.

As previously stated, a speedier throw-in has a higher probability of being successful and retaining possession. Despite this, there is an undeniable upward trend in the amount of time spent on throw-ins since the preceding action (see in Figure 23). In general, longer throw-ins, on average, are less likely to be completed and possession retained within seven seconds. As shown in Figure 24, the average distance decreases over the last seasons; however, the pattern in the attacking zone is not as obvious as it is in other defined zones; nevertheless, this is not a problem, as we have learned, long throw-ins aimed into the penalty area pose a greater danger to opponents.

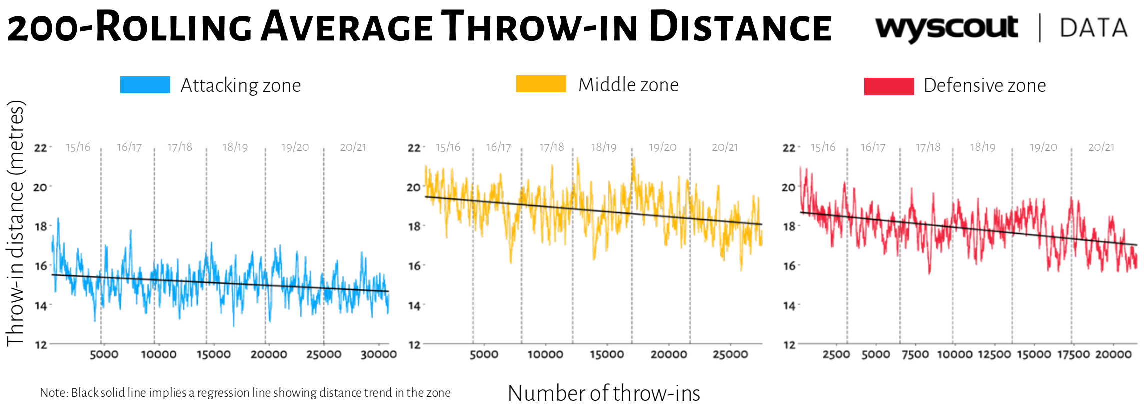
Figure 23: 200-rolling average of the throw-in timing



Note: Black solid line implies a regression line showing timing trend in the zone

Source: own elaboration from Wyscout Data

Figure 24: 200-rolling average of the throw-in distance

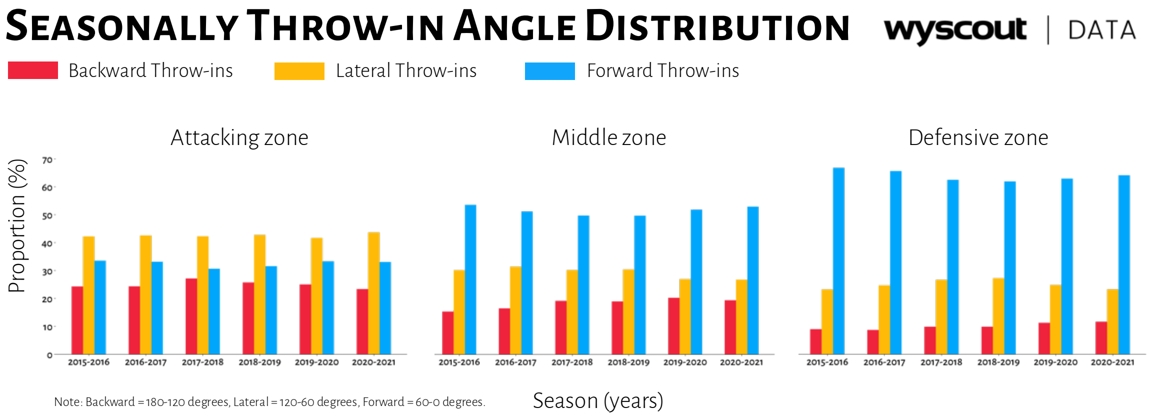


Note: Black solid line implies a regression line showing distance trend in the zone

Source: own elaboration from Wyscout Data

To improve readability, we created bar charts to depict the patterns in the distribution of throw-in angles. As described in Section 3.3., we classified throw-in angles into three categories: backward throw-ins, lateral throw-ins, and forward throw-ins. As with the previous two line-plots, the bar chart (in Figure 25) was divided into three distinct zones to help in analysing trends. In the defensive and middle zone, there has been a decrease in the percentage of lateral throw-ins and a small increase in the percentage of forward throw-ins since season 2017/2018, which might be linked to fear of losing control closer to their own goal. Since season 2017/2018, there has been a minor reduction in the proportion of backwards throw-ins and a slight increase in the proportion of lateral throw-ins in the attacking zone. Although this trend is not significant, it is a good start toward creating danger from attacking throw-ins, because, as said before, more progressive throw-ins increase the probability of scoring a goal and avoiding a conceded goal, which is supposed to be the objective, primarily in the attacking zone.

Figure 25: The distribution of the throw-in angle of each season since 2015/2016

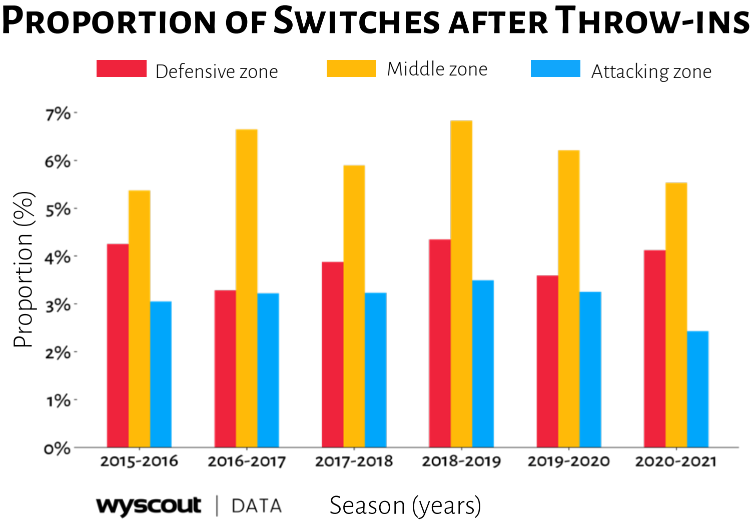


Note: Backward = 180-120 degrees, Lateral = 120-60 degrees, Forward = 60-0 degrees.

Source: own elaboration from Wyscout Data

For the middle and defensive zones, switches are an effective way to prevent allowing a goal whilst also simultaneously creating danger for the opponent by creating a valuable scoring opportunity. Switches after throw-ins within 7 s is also worth noting, since there has been a positive drop in switches following attacking throw-ins since season 2017/2018, which is ineffective in terms of xGD, as we previously examined. With regard to xGD, switches after throw-ins from the centre zone (as well as the defensive zone) are an effective weapon. Although this is the case, there has been a notable decrease in the proportion of switches after throw-ins that started in the middle zone. A bar-plot was produced to represent all of these patterns in the proportion of throw-ins with a switch within 7 s, separated into three different zones, with a legend describing which colour corresponds to which zone (see Figure 26).

Figure 26: The proportion of throw-ins with a switch for each season since 2015/2016



Source: own elaboration from Wyscout Data

4.2 Expected Throw model

After gaining insight into how each variable affects the target variable (completion), as well as other statistical indicators (such as possession retention and xG), we can begin building a machine learning model using these characteristics. The target variable predicts whether a throw-in targets a teammate based on the following attributes inspired by an existing machine learning model constructed by Eliot McKinley (McKinley, 2018), who trained the model on almost identical features on Major League Soccer data without benchmarking. Thus, our aim is to analyse how each variable affects our model using data from the Czech Football League and to determine how accurate the model is, using events data from Wyscout (Wyscout, 2022).

4.2.1 Feature Engineering and Model Selection

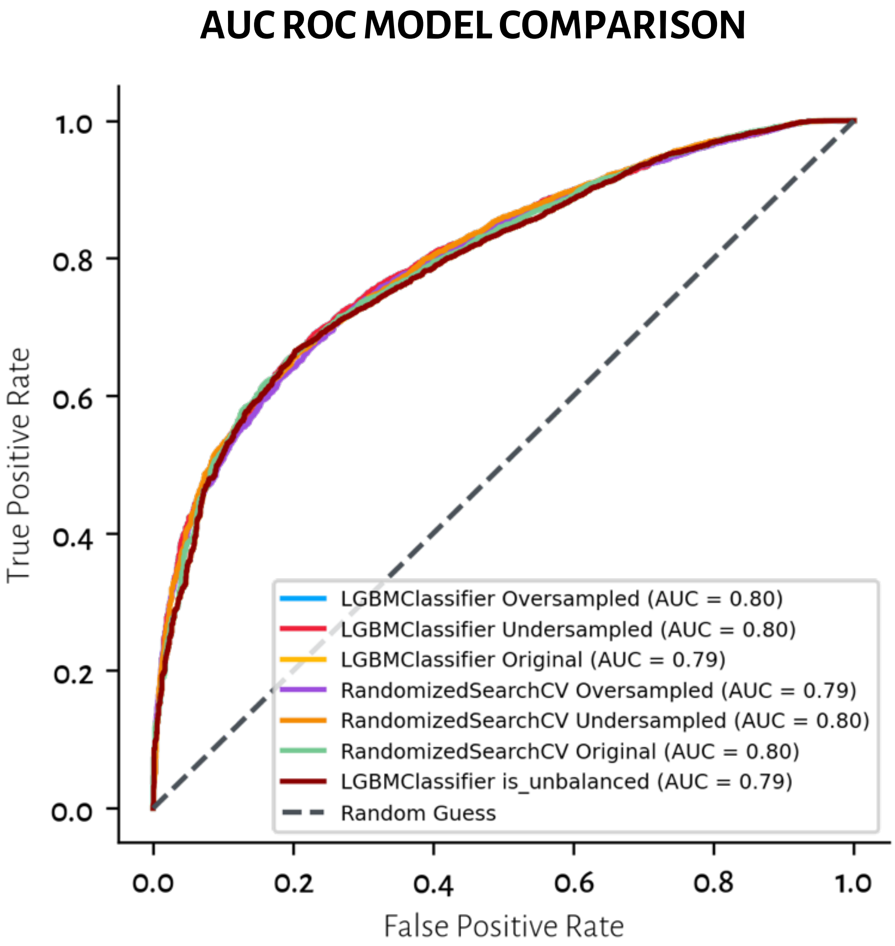
We have an unbalanced dataset, where a negative class (unsuccessful throw-ins) is represented by 10 % of the whole dataset, while a positive class (successful throw-ins) is represented by 90 % of the whole dataset. Therefore, we compare our dataset to two synthetic datasets. The first dataset is unmodified, the second is over-sampled version of our dataset, and the third is under-sampled version of our dataset. Then, we need to split each dataset to train, validation and test sets. Thus, we also must modify categorical features in order to use over- and under-sampling. As a result, a game state is divided into two dummy features: 'game_state_is_leading' and 'game_state_is_losing'. This is accomplished through the use of one-hot encoding, a common approach for dealing with categorical data.

4.2.2 Fitting model

Before fitting the model, we use the RS method (see Section 2.5) to find the best hyperparameters for model learning. We train the LGBM Classifier model twice on each dataset (unmodified, over-sampled, and under-sampled) and prepare them for comparison. To summarize, the first cycle fits the model using default parameters, whereas the second round fits the model using a unique combination of hyper-parameters discovered by the RS method. We analyse the ROC and PR curves of unseen sets to determine which one to utilise for prediction. ROC plot in Figure 27 shows that any models with modified datasets are not better in contrast to the LGBM Classifier model with RS hyper-parameters of the original dataset with an AUC of 0.80 (see green colour in Figure 27).

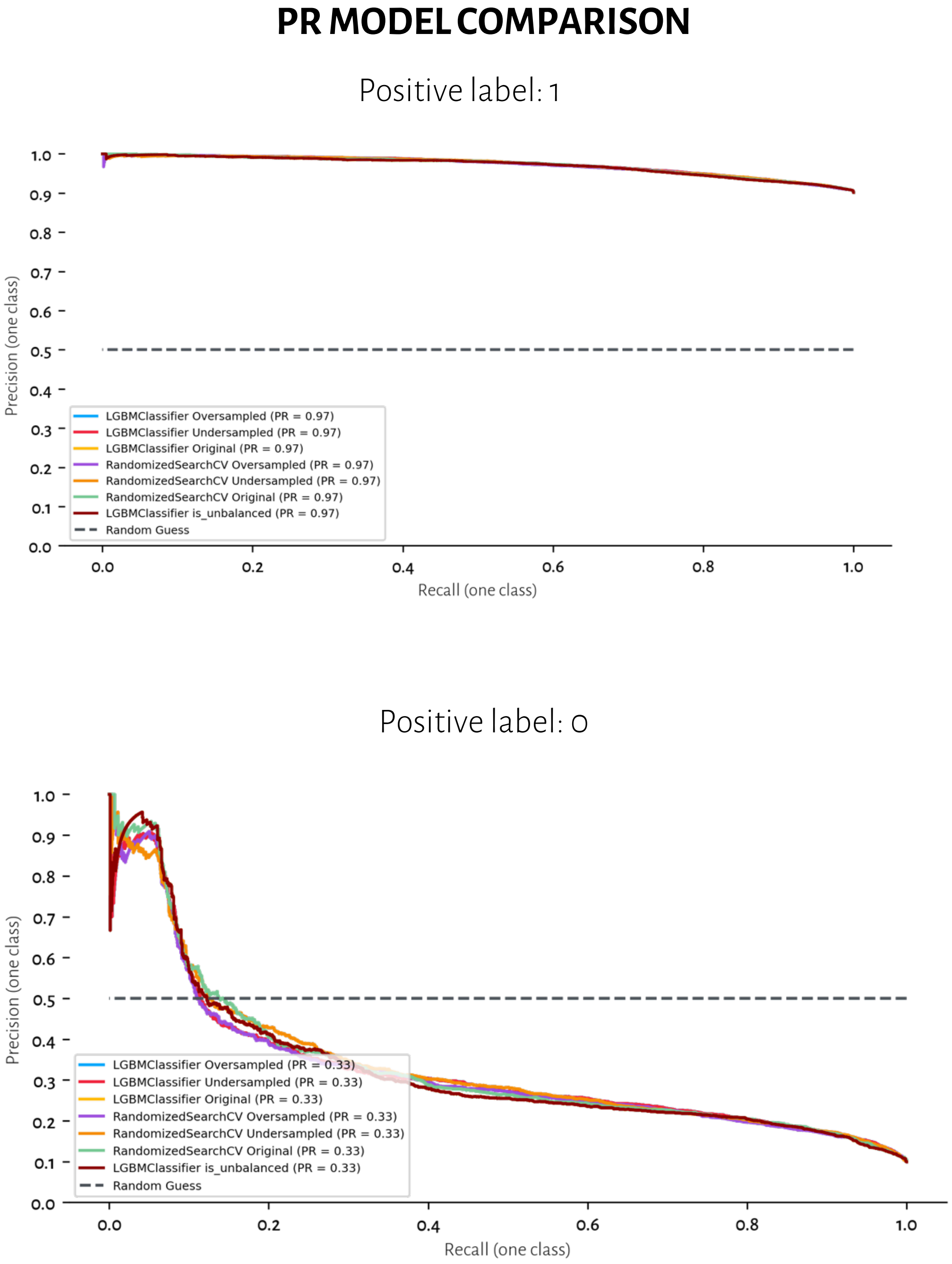
Figure 28 compares models based on the relationship between precision and recall for each class separately and the PR plot also shows any model with modified datasets are not better in contrast to the LGBM Classifier model with RS hyper-parameters of the original dataset, which has an PR of 0.97 and 0.33 for the positive and the negative label respectively. As a conclusion, any models with modified datasets seem not to be better in contrast the LGBMClassifier model with the original dataset.

Figure 27: AUC ROC Model comparison



Source: own elaboration

Figure 28: Precision-Recall Model comparison



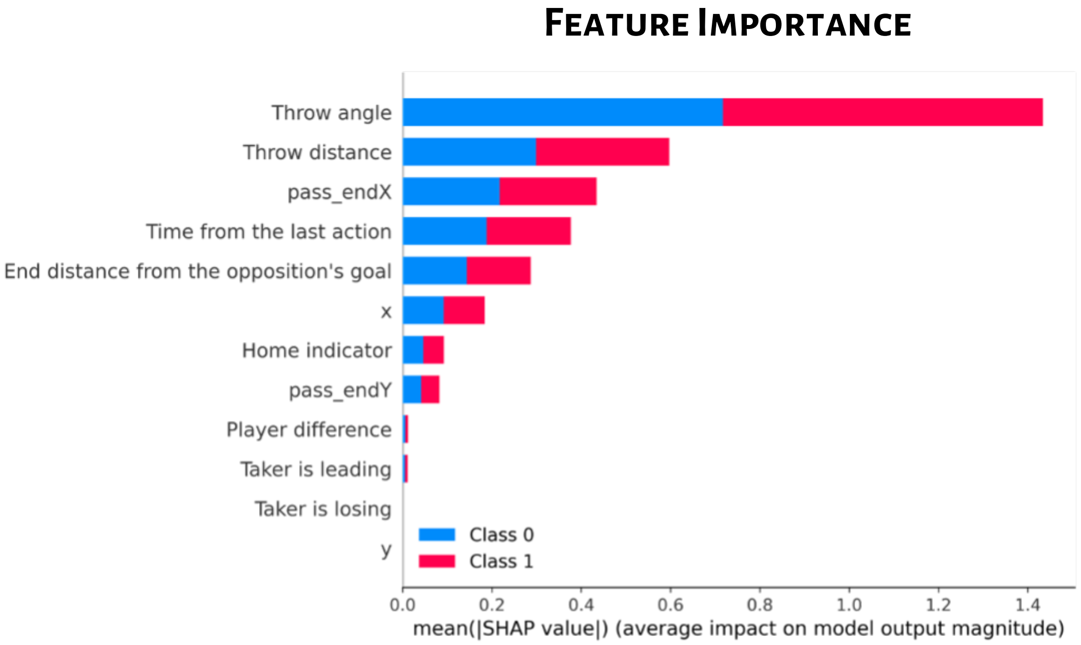
Source: own elaboration

4.2.3 Model Explanation

To extract the greatest value from the model, we use the SHAP plots (Lundberg, 2018)to evaluate the model's feature importance and the effect of each feature on the prediction of each instance. Figure 29 displays a bar plot with a feature importance of our model. The throw-angle is the most critical element, the least important is the y-coordinate of a throw-in, which determines which side of throw-ins are being taken. The relationship between each attribute and its impact on the prediction is shown in Figure 30.

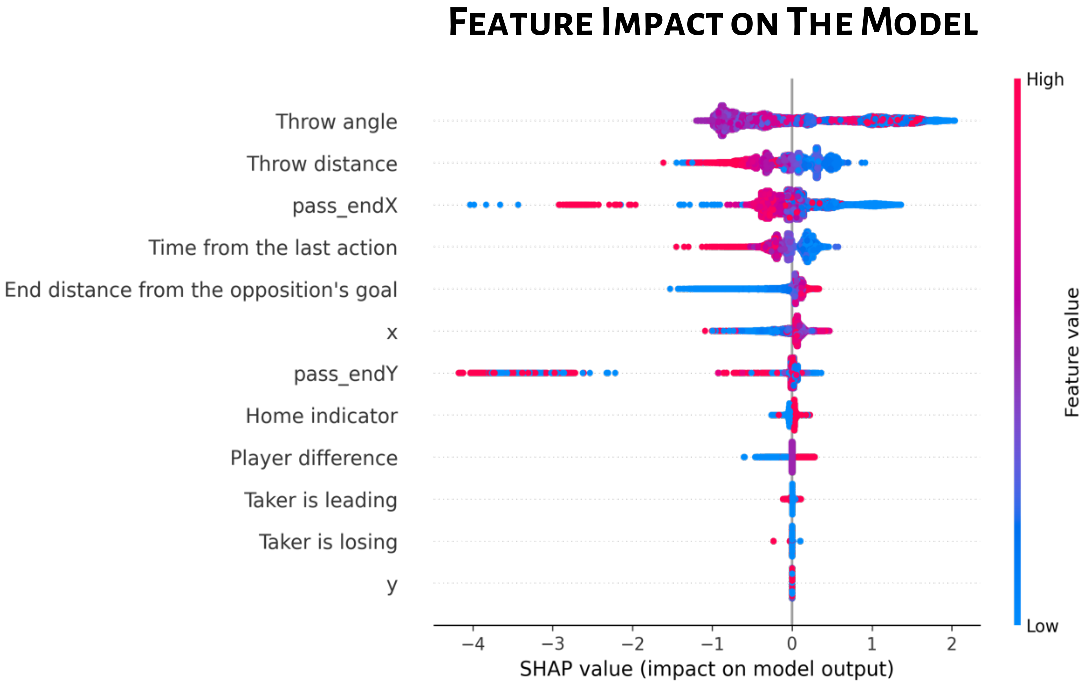
As seen in Figure 30, Throw distance reflects our pre-existing relation to completion; for throw-ins with a longer distance, the negative SHAP value outweighs the positive SHAP value, indicating that throw-ins with a greater distance are more likely to be incomplete. Time from the previous action is equivalent to the throw-in distance in terms of the effect on the target variable; moreover, the positive value outweighs the negative value for a faster throw-in, meaning that the quicker throw-ins are more likely to be completed. On the other hand, the negative value outweighs the positive value for throw-ins taken by a team with one less player and for throw-ins with a closer end distance to the opposing goal. All of these connections are consistent with the probabilities we evaluated in EDA.

Figure 29: Feature importance (using the SHAP value)



Source: own elaboration using library from (Lundberg, 2018)

Figure 30: Feature impact on the model output (using the SHAP value)



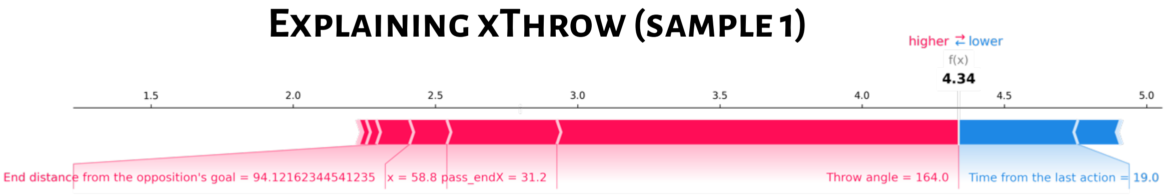
Source: own elaboration using library from (Lundberg, 2018)

We may get insight into how models form decisions by examining SHAP plots. The presence of red bars means that the risk of failing to complete the throw-in is decreasing, while the blue bars shows that the risk of failing to complete the throw-in is increasing. The bar length of SHAP force plot (Lundberg, 2018) indicates the significance of a certain feature in predicting completion of a particular throw-in. The least important features for predicting completion of a certain throw-in are not displayed in the force plot. Base value indicates the average.

For illustration, Figure 31 indicates that a relatively backward angle (throw angle=164) greatly minimizes the likelihood of a failed throw-in. Additionally, starting in the middle zone (x=58.8) reduces the likelihood of a missed throw-in. Whereas the amount of time passed since the last action increases the risk that the throw-in will be incomplete. A high output value (4.34) above the SHAP force plot indicates that the throw-in is very likely to reach the teammate based on trained features.

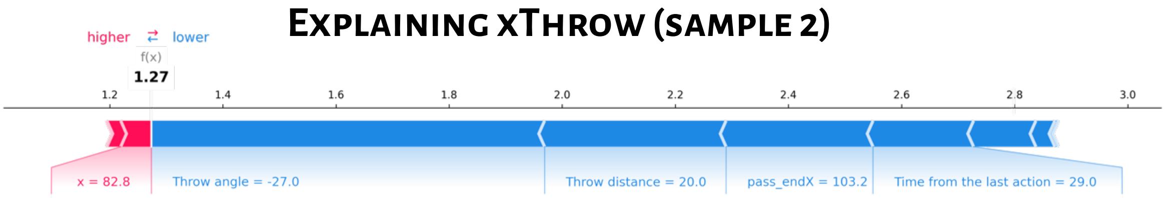
On the other hand, Figure 32 indicates that a pretty forward angle (throw angle=-27) greatly increase the likelihood of a failed throw-in. Throw-in distance (throw distance=20) relatively increase the risk of an unsuccessful throw-in. Ending position of the throw-in in the middle zone (pass_endX=103) reduces the likelihood of a successful throw-in. On the other side, the amount of time passed since the last action increases the risk that the throw-in will be incomplete.

Figure 31: The model explanation on a throw-in as a sample 1



Source: own elaboration using SHAP Force plot from (Lundberg, 2018)

Figure 32: The model explanation on a throw-in as a sample 2



Source: own elaboration using SHAP Force plot from (Lundberg, 2018)

5 Conclusion

The goal of this thesis was to construct a prediction model for throw-in completion using the machine learning framework LightGBM in order to acquire a better knowledge of the characteristics of throw-ins that impact the completion rate. While the original dataset was split into training data (80%) and testing data (20%) for the purpose of evaluating the model using ROC and PR curves, for the over-sampling and under-sampling datasets, a validation dataset was created by splitting the training dataset into a smaller training dataset and a new validation dataset to evaluate the model performance trained on over-sampling and under-sampling datasets. While classification of successful throw-ins performs well, classification of failed throw-ins performs poorly, despite the fact that we created over- and under-sampling datasets to balance the proportion of those two classes, or even if we used the RS to find the optimal set of model hyperparameters. As we predicted prior to the practical section, we assumed that performance would be less than optimal due to the fact that we do not have an adequate amount of features to develop an optimal model. However, we believed that we accomplished the objective of the thesis, as we were able to interpret how each characteristic influences the target variable and other statistical indicators associated with the assessment of the throw-ins.

Additionally, the second purpose was to assess the effect of the characteristics of the throw-in on other statistical indicators such as possession retention and scoring probability via xG. We proved based on xG and Gls statistics that attacking throw-ins aimed into the penalty box increases the chance of scoring and avoiding a conceded goal in contrast to attacking throw-ins aimed outside of the penalty box. The results also suggests that it is beneficial to wait for the optimal time; from domain knowledge, we may argue that it is advantageous to wait until players have formed in their positions, and also the taker has wiped the ball for a stronger grip. Generally, the faster throw-ins are likely to be successful and possession held. The shorter throw-ins are likely to be successful and possession retained. Backward throw-ins have higher chance of being successful and possession kept in comparison to forward throw-ins. The leading teams and teams with one extra player often take less progressive throw-ins. Lastly, we studied that a throw-in with a switch, originating in the defensive zone or middle zone, has generally a better likelihood of scoring a goal and also avoiding a conceded goal based on xGD metrics.

6 References

Soccerment Research, 2021. *The Growing Importance of Football Analytics - Soccerment Research.* [Online] Available at: https://soccerment.com/the-importance-of-football-analytics/ [Accessed December 2021].

Wyscout, 2022. *Make the most of Wyscout Data.* [Online] Available at: https://footballdata.wyscout.com [Accessed January 2022].

Wyscout Spa, 2022. *The Football Platform.* [Online] Available at: https://wyscout.com [Accessed December 2021].

Wyscout, 2021. *xG - Wyscout Glossary.* [Online] Available at: https://dataglossary.wyscout.com/xg/ [Accessed December 2021].

Whitmore, J., 2021. *What Are Expected Goals (xG)? - The Analyst.* [Online] Available at: https://theanalyst.com/eu/2021/07/what-are-expected-goals-xg/ [Accessed January 2022].

Kullowatz, M., 2018. *An Updated Expected Passing Model — American Soccer Analysis.* [Online] Available at: https://www.americansocceranalysis.com/home/2018/4/19/an-updated-expected-passing-model [Accessed January 2022].

McKinley, E., 2018. *Game of Throw-ins — American Soccer Analysis.* [Online] Available at: https://www.americansocceranalysis.com/home/2018/11/27/game-of-throw-ins [Accessed January 2022].

Stone, J. A., Smith, A. & Barry, A., 2021. The undervalued set piece: Analysis of soccer throw-ins during the English Premier League 2018–2019 season. *International Journal of Sports Science & Coaching,* June, 16(3), pp. 830-839.

Guinness World Records, 2019. *Farthest distance football (soccer) throw-in (male).* [Online] Available at: https://www.guinnessworldrecords.com/world-records/longest-throw-in-(football) [Accessed January 2022].

International Football Association Board, 2021. *LAW OF THE GAME - Football Laws, Rules and Regulations.* [Online] Available at: https://www.theifab.com [Accessed January 2022].

Goos, P. & Meintrup, D., 2015. *Statistics with JMP: Graphs, Descriptive Statistics and Probability,* New York: John Wiley et Sons Inc.

Aggarwal, C. C. & Reddy, C. K., 2013. *Data Clustering : Algorithms and Applications.* CK (eds) 2013 ed. Philadelphia: CRC Press LLC.

DeepAI, 2022. *Hyperparameters - DeepAI.* [Online] Available at: https://deepai.org/machine-learning-glossary-and-terms/hyperparameter [Accessed February 2022].

Bergstra, J. & Bengio, Y., 2012. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.,* pp. 281-305.

Davis, J. & Goadrich, M., 2006. The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning - ICML '06,* pp. 233-240.

Ke, G. et al., 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems 30,* p. 3146–3154.

Gan, M. et al., 2021. Application of the Machine Learning LightGBM Model to the Prediction of the Water Levels of the Lower Columbia River. *Journal of Marine Science and Engineering,* 9(5), pp. 0-22.

Microsoft Corporation, 2022. *Features - LightGBM.* [Online] Available at: https://lightgbm.readthedocs.io/en/latest/Features.html [Accessed January 2022].

Microsoft Corporation, 2022. *LightGBM, Release 3.3.2.99.* [Online] Available at: https://buildmedia.readthedocs.org/media/pdf/lightgbm/latest/lightgbm.pdf [Accessed March 2022].

Microsoft Corporation, 2022. *Microsoft.* [Online] Available at: https://www.microsoft.com [Accessed January 2022].

Lundberg, S. M. & Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions. *Curran Associates Inc.,* 24 November.pp. 1-10.

Lundberg, S., 2018. *Welcome to the SHAP documentation.* [Online] Available at: https://shap.readthedocs.io/en/latest/index.html [Accessed February 2022].

7 List of figure, tables, and abbreviations

7.1 List of figures

Figure 1: The Throw-in Distance Box Plot........12
Figure 2: Types of Tree Growth........16
Figure 3: Flow diagram of optimizing set of hyperparameters........17
Figure 4: Frequently Used Machine Learning Evaluation Metrics........18
Figure 5: Throw-ins Glossary........25
Figure 6: Completion and Possession Retention Rate of Throw-ins by the End Locations 28
Figure 7: Throwing Sonars........29
Figure 8: xGD from throw-ins by the end location........30
Figure 9: Probabilities of throw-ins by the throw-in distance........31
Figure 10: The distribution of long attacking throw-ins........32
Figure 11: Probabilities of throw-ins by the throw-in angle with sonars........33
Figure 12: Probabilities of throw-ins by timing........34
Figure 13: Defensive throw-ins combos........36
Figure 14: Attacking throw-ins combos........36
Figure 15: xGD with / without a switch after a throw-in........38
Figure 16: xGA from Throw-in's Turnovers........39
Figure 17: Probabilities of attacking throw-ins by the end location........41
Figure 18: xGD from attacking throw-ins by timing........41
Figure 19: Probabilities of throw-ins by the game state........42
Figure 20: The distribution of throw-ins by the game state........44
Figure 21: Probabilities of throw-ins by the player difference........45
Figure 22: The distribution of throw-ins by the player difference........46
Figure 23: 200-rolling average of the throw-in timing........48
Figure 24: 200-rolling average of the throw-in distance........48
Figure 25: The distribution of the throw-in angle of each season since 2015/2016........49
Figure 26: The proportion of throw-ins with a switch for each season since 2015/2016....50
Figure 27: AUC ROC Model comparison........52
Figure 28: Precision-Recall Model comparison........53
Figure 29: Feature importance (using the SHAP value)........54

Figure 30: Feature impact on the model output (using the SHAP value) 55
Figure 31: The model explanation on a throw-in as a sample 1 56
Figure 32: The model explanation on a throw-in as a sample 2 56

7.2 List of tables

Table 1: Trained features by the LightGBM model 13
Table 2: The model's primary hyperparameters employed in this research 17
Table 3: Probabilities of Types of Throw-in 27
Table 4: Attacking throw-ins xGD and GD comparison table 40
Table 5: Feature Summary 47

7.3 List of abbreviations

xG: Expected Goals

xGA: Expected Goals against

xGD: Expected Goals difference

Gls: Goals

GD: Goal difference

xPass: Expected Pass

xThrow: Expected Throw

Combos: Passing Combinations

RS: Random Search

GS: Grid Search

SHAP: **SHapley Additive exPlanation**

ROC: Receiver Operating Characteristic

PR: Precision-Recall

TRP: The True Positive Rate

FPR: The False Positive Rate

s: second(s)

mts: metre(s)